

**This Page is Inserted by IFW Indexing and Scanning
Operations and is not part of the Official Record**

BEST AVAILABLE IMAGES

Defective images within this document are accurate representations of the original documents submitted by the applicant.

Defects in the images include but are not limited to the items checked:

- BLACK BORDERS**
- IMAGE CUT OFF AT TOP, BOTTOM OR SIDES**
- FADED TEXT OR DRAWING**
- BLURRED OR ILLEGIBLE TEXT OR DRAWING**
- SKEWED/SLANTED IMAGES**
- COLOR OR BLACK AND WHITE PHOTOGRAPHS**
- GRAY SCALE DOCUMENTS**
- LINES OR MARKS ON ORIGINAL DOCUMENT**
- REFERENCE(S) OR EXHIBIT(S) SUBMITTED ARE POOR QUALITY**
- OTHER:** _____

IMAGES ARE BEST AVAILABLE COPY.

As rescanning these documents will not correct the image problems checked, please do not report these problems to the IFW Image Problem Mailbox.

09/4/2, 249

日本国特許庁
PATENT OFFICE
JAPANESE GOVERNMENT

別紙添付の書類に記載されている事項は下記の出願書類に記載されて
いる事項と同一であることを証明する。

This is to certify that the annexed is a true copy of the following application as filed
with this Office.

出願年月日
Date of Application:

1999年12月 2日

出願番号
Application Number:

平成11年特許願第343890号

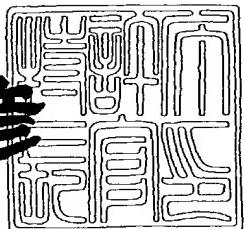
出願人
Applicant(s):

株式会社リコー

2000年 1月 7日

特許庁長官
Commissioner,
Patent Office

近藤 隆彦



出証番号 出証特平11-3091726

【書類名】 特許願

【整理番号】 9907814

【提出日】 平成11年12月 2日

【あて先】 特許庁長官殿

【国際特許分類】 G06F 15/40
G06F 17/27
G06F 3/00

【発明の名称】 文書処理装置、文書分類装置、文書処理方法、文書分類方法およびそれらの方法をコンピュータに実行させるプログラムを記録したコンピュータ読み取り可能な記録媒体

【請求項の数】 38

【発明者】

【住所又は居所】 東京都大田区中馬込1丁目3番6号 株式会社リコー内

【氏名】 嶋田 敦夫

【発明者】

【住所又は居所】 東京都大田区中馬込1丁目3番6号 株式会社リコー内

【氏名】 宮地 達生

【発明者】

【住所又は居所】 東京都大田区中馬込1丁目3番6号 株式会社リコー内

【氏名】 劍持 栄治

【発明者】

【住所又は居所】 東京都大田区中馬込1丁目3番6号 株式会社リコー内

【氏名】 山崎 真湖人

【発明者】

【住所又は居所】 東京都大田区中馬込1丁目3番6号 株式会社リコー内

【氏名】 武谷 一寿

【発明者】

【住所又は居所】 東京都大田区中馬込1丁目3番6号 株式会社リコー内

【氏名】 長束 哲郎

【特許出願人】

【識別番号】 000006747

【氏名又は名称】 株式会社リコー

【代理人】

【識別番号】 100104190

【弁理士】

【氏名又は名称】 酒井 昭徳

【先の出願に基づく優先権主張】

【出願番号】 平成10年特許願第376576号

【出願日】 平成10年12月24日

【先の出願に基づく優先権主張】

【出願番号】 平成10年特許願第369589号

【出願日】 平成10年12月25日

【先の出願に基づく優先権主張】

【出願番号】 平成11年特許願第 22915号

【出願日】 平成11年 1月29日

【手数料の表示】

【予納台帳番号】 041759

【納付金額】 21,000円

【提出物件の目録】

【物件名】 明細書 1

【物件名】 図面 1

【物件名】 要約書 1

【包括委任状番号】 9810808

【ブルーフの要否】 要

【書類名】 明細書

【発明の名称】 文書処理装置、文書分類装置、文書処理方法、文書分類方法およびそれらの方法をコンピュータに実行させるプログラムを記録したコンピュータ読み取り可能な記録媒体

【特許請求の範囲】

【請求項1】 入力された複数の文書データを所定の形式で表示または印刷するために出力する文書処理装置において、

入力された文書データを記憶する文書記憶手段と、

前記文書記憶手段により記憶された文書データの全部または一部を選択する選択手段と、

前記選択手段により選択された文書データの全部または一部から文字列の特徴に関するデータを抽出する特徴抽出手段と、

前記特徴抽出手段により抽出された文字列の特徴に関するデータに基づいて前記文書データの全部または一部を加工処理する加工処理手段と、

前記加工処理手段により加工処理された文書データの全部または一部を出力する出力手段と、

を備えたことを特徴とする文書処理装置。

【請求項2】 前記出力手段は、

前記加工処理手段により加工処理された文書データの全部または一部の内容に基づいて複数の項目値を設定する項目値設定手段と、

前記項目値設定手段により設定された項目値ごとに前記文書データの全部または一部を集計する集計手段と、

を備え、

前記文書データの全部または一部を、項目値を少なくとも一つの軸とする表形式に展開して出力することを特徴とする請求項1に記載の文書処理装置。

【請求項3】 前記出力手段は、さらに、前記加工処理手段により加工処理された文書データの全部または一部を、前記加工処理手段により加工処理される前の文書データの全部または一部とともに出力することを特徴とする請求項1または2に記載の文書処理装置。

【請求項4】 前記文書記憶手段は、さらに、前記加工処理手段により加工処理された文書データの全部または一部を記憶することを特徴とする請求項1～3のいずれか一つに記載の文書処理装置。

【請求項5】 前記選択手段は、さらに、前記出力手段により出力された文書データの全部または一部を選択することを特徴とする請求項1～4のいずれか一つに記載の文書処理装置。

【請求項6】 前記文書記憶手段は、さらに、前記加工処理の内容に関するデータを記憶することを特徴とする請求項1～5のいずれか一つに記載の文書処理装置。

【請求項7】 文書の内容に基づいて文書の分類をおこなう文書分類装置において、

文書データを入力する入力手段と、

前記入力手段により入力された文書データを解析して言語解析情報を得る言語解析手段と、

前記言語解析手段により得られた言語解析情報に基づいて前記文書データに対する文書特徴ベクトルを生成するベクトル生成手段と、

前記ベクトル生成手段により生成された文書特徴ベクトル間の類似度に基づいて文書を分類し、文書の部分集合を生成する分類手段と、

前記分類手段により生成された文書の部分集合の特徴であるクラスタ特徴を算出するクラスタ特徴算出手段と、

前記クラスタ特徴算出手段により算出されたクラスタ特徴を分類体系の構成要素として記憶する分類体系記憶手段と、

を備えたことを特徴とする文書分類装置。

【請求項8】 文書の内容に基づいて文書の分類をおこなう文書分類装置において、

文書データを入力する入力手段と、

前記入力手段により入力された文書データを解析して言語解析情報を得る言語解析手段と、

前記言語解析手段により得られた言語解析情報に基づいて前記文書データに対

する文書特徴ベクトルを生成するベクトル生成手段と、

前記ベクトル生成手段により生成された文書特徴ベクトル間の類似度に基づいて文書を分類し、文書の部分集合を生成する分類手段と、

前記分類手段により生成された文書の部分集合の特徴であるクラスタ特徴を算出するクラスタ特徴算出手段と、

前記クラスタ特徴算出手段により算出されたクラスタ特徴を表示する表示手段と、

前記分類手段により生成された文書の部分集合の中から所望の部分集合を選択するクラスタ選択指示手段と、

前記クラスタ選択指示手段により選択された文書の部分集合を分類体系の構成要素として記憶する分類体系記憶手段と、

を備えたことを特徴とする文書分類装置。

【請求項9】 前記ベクトル生成手段により生成された文書特徴ベクトルを記憶する文書特徴ベクトル記憶手段と、

前記文書特徴ベクトル記憶手段により記憶された文書特徴ベクトルを、前記クラスタ選択指示手段により選択された部分集合に属する文書の文書特徴ベクトルを除去したのこりとなるように修正するベクトル修正手段と、

を備え、

前記分類手段は、前記ベクトル修正手段により修正された文書特徴ベクトルに基づいて文書を分類することを特徴とする請求項8に記載の文書分類装置。

【請求項10】 前記ベクトル生成手段により生成された文書特徴ベクトルを記憶する文書特徴ベクトル記憶手段と、

前記文書特徴ベクトル記憶手段により記憶された文書特徴ベクトル間の類似度を判断する際の文書表現空間を前記クラスタ選択指示手段により選択された部分集合から算出する特徴量に基づいて修正する文書表現空間修正手段と、

を備え、

前記分類手段は、前記文書表現空間修正手段により修正された文書表現空間を用いて、前記ベクトル生成手段により生成された文書特徴ベクトル間の類似度に基づいて文書を分類することを特徴とする請求項8に記載の文書分類装置。

【請求項11】 前記ベクトル生成手段により生成された文書特徴ベクトルを記憶する文書特徴ベクトル記憶手段と、

前記文書特徴ベクトル記憶手段により記憶された文書特徴ベクトル間の類似度を判断する際の文書表現空間を前記クラスタ選択指示手段により選択された部分集合から算出する特微量に基づいて修正する文書表現空間修正手段と、

を備え、

前記分類手段は、前記文書表現空間修正手段により修正された文書表現空間を用いて、前記ベクトル修正手段により修正された文書特徴ベクトル間の類似度に基づいて文書を分類することを特徴とする請求項9に記載の文書分類装置。

【請求項12】 前記分類手段により生成された文書の部分集合に所属する文書のすべてあるいは一部が選択された場合に選択されたことを示す選択情報を付与する選択情報付与手段を備え、

前記表示手段は、前記クラスタ特徴を表示するとともに、前記選択情報付与手段により付与された選択情報を表示することを特徴とする請求項8または10に記載の文書分類装置。

【請求項13】 前記分類体系記憶手段は、前記選択指示手段により選択された文書の部分集合に属する全部あるいは一部の文書のほか、クラスタ特徴および／または操作者が作成した任意の情報を分類体系の構成要素として記憶することを特徴とする請求項8～12に記載の文書分類装置。

【請求項14】 文書の内容にしたがって文書群を分類する文書分類装置において、

文書データ群を入力する文書入力手段と、

入力された文書データ群の各文書に対して所定の基準に基づき文書の分割をおこない、一つの文書データから一つまたは複数の分割文書データを生成する文書分割手段と、

前記文書データと前記分割文書データとの対応を示す文書一分割文書対応マップを生成する文書一分割文書対応マップ生成手段と、

前記分割文書データを分類する分割文書分類手段と、

前記分割文書分類手段による分類結果に基づいて分割文書分類結果情報を生成

する分割文書分類結果生成手段と、

前記文書一分割文書対応マップと前記分割文書分類結果情報とを用いて前記文書データの分類結果情報を生成する文書分類結果生成手段と、

を備えたことを特徴とする文書分類装置。

【請求項15】 前記文書データを保存する文書保存手段と、
前記分割文書データを保存する分割文書保存手段と、
前記文書一分割文書対応マップ生成手段により生成された文書一分割文書対応マップを保存する文書一分割文書対応マップ保存手段と、
を備えたことを特徴とする請求項14に記載の文書分類装置。

【請求項16】 前記分割文書分類結果生成手段により生成された分割文書分類結果情報を保存する分割文書分類結果保存手段を備えたことを特徴とする請求項15に記載の文書分類装置。

【請求項17】 前記文書分割手段により生成される複数の分割文書データには分割前の文書データそのものを含むことを特徴とする請求項14～16のいずれか一つに記載の文書分類装置。

【請求項18】 前記文書分割手段が、文書データの構造情報を基に文書データを分割する構成にしたことを特徴とする請求項14～17のいずれか一つに記載の文書分類装置。

【請求項19】 前記文書データに含まれる要素を抽出する文書要素抽出手段と、

前記文書要素抽出手段により抽出された要素に付随する要素付隨情報を抽出する要素付隨情報抽出手段と、

を備え、

前記文書分割手段が、前記文書要素抽出手段により抽出された要素、または前記要素と前記要素付隨情報抽出手段により抽出された要素付隨情報を用いて前記文書データを分割する構成にしたことを特徴とする請求項14～17のいずれか一つに記載の文書分類装置。

【請求項20】 前記文書分割手段が、指示された指定範囲にしたがって文書データの分割をおこなう構成にしたことを特徴とする請求項14～17のいず

れか一つに記載の文書分類装置。

【請求項21】 前記文書分割手段が、文書データ中の文字数、文数、または文字数と文数の両方を基に文書データを分割する構成にしたことを特徴とする請求項14～17のいずれか一つに記載の文書分類装置。

【請求項22】 前記文書分類結果生成手段が、文書データを示す情報および前記文書データに付随する代表的情報を、分類結果情報として抽出して提示する構成にしたことを特徴とする請求項14～21のいずれか一つに記載の文書分類装置。

【請求項23】 前記文書分類結果生成手段が、分割文書データを示す情報および前記分割文書データに付随する代表的情報を、分類結果情報として、抽出して提示する構成にしたことを特徴とする請求項22に記載の文書分類装置。

【請求項24】 入力された複数の文書データを所定の形式で表示または印刷するために出力する文書処理方法において、

入力された文書データを記憶する文書記憶工程と、

前記文書記憶工程により記憶された文書データの全部または一部を選択する選択工程と、

前記選択工程により選択された文書データの全部または一部から文字列の特徴に関するデータを抽出する特徴抽出工程と、

前記特徴抽出工程により抽出された文字列の特徴に関するデータに基づいて前記文書データの全部または一部を加工処理する加工処理工程と、

前記加工処理工程により加工処理された文書データの全部または一部を出力する出力工程と、

を含んだことを特徴とする文書処理方法。

【請求項25】 前記出力工程は、

前記加工処理工程により加工処理された文書データの全部または一部の内容に基づいて複数の項目値を設定する項目値設定工程と、

前記項目値設定工程により設定された項目値ごとに前記文書データの全部または一部を集計する集計工程と、

を含み、

前記文書データの全部または一部を、項目値を少なくとも一つの軸とする表形式に展開して出力することを特徴とする請求項24に記載の文書処理方法。

【請求項26】 前記出力工程は、さらに、前記加工処理工程により加工処理された文書データの全部または一部を、前記加工処理工程により加工処理される前の文書データの全部または一部とともに出力することを特徴とする請求項24または25に記載の文書処理方法。

【請求項27】 前記文書記憶工程は、さらに、前記加工処理工程により加工処理された文書データの全部または一部を記憶することを特徴とする請求項24～26のいずれか一つに記載の文書処理方法。

【請求項28】 前記選択工程は、さらに、前記出力工程により出力された文書データの全部または一部を選択することを特徴とする請求項24～27のいずれか一つに記載の文書処理方法。

【請求項29】 前記文書記憶工程は、さらに、前記加工処理の内容に関するデータを記憶することを特徴とする請求項24～28のいずれか一つに記載の文書処理方法。

【請求項30】 文書の内容に基づいて文書の分類をおこなう文書分類方法において、

文書データを入力する入力工程と、
前記入力工程により入力された文書データを解析して言語解析情報を得る言語解析工程と、

前記言語解析工程により得られた言語解析情報に基づいて前記文書データに対する文書特徴ベクトルを生成するベクトル生成工程と、

前記ベクトル生成工程により生成された文書特徴ベクトル間の類似度に基づいて文書を分類し、文書の部分集合を生成する分類工程と、

前記分類工程により生成された文書の部分集合の特徴であるクラスタ特徴を算出するクラスタ特徴算出工程と、

前記クラスタ特徴算出工程により算出されたクラスタ特徴に基づいて分類体系の構成要素を生成する分類体系生成工程と、
を含んだことを特徴とする文書分類方法。

【請求項31】 文書の内容に基づいて文書の分類をおこなう文書分類方法において、

文書データを入力する入力工程と、

前記入力工程により入力された文書データを解析して言語解析情報を得る言語解析工程と、

前記言語解析工程により得られた言語解析情報に基づいて前記文書データに対する文書特徴ベクトルを生成するベクトル生成工程と、

前記ベクトル生成工程により生成された文書特徴ベクトル間の類似度に基づいて文書を分類し、文書の部分集合を生成する分類工程と、

前記分類工程により生成された文書の部分集合の特徴であるクラスタ特徴を算出するクラスタ特徴算出工程と、

前記クラスタ特徴算出工程により算出されたクラスタ特徴を表示する表示工程と、

前記分類工程により生成された文書の部分集合の中から所望の部分集合を選択するクラスタ選択指示工程と、

前記クラスタ選択指示工程により選択されたクラスタ特徴に基づいて分類体系の構成要素を生成する分類体系生成工程と、

を含んだことを特徴とする文書分類方法。

【請求項32】 前記ベクトル生成工程により生成された文書特徴ベクトルを、前記クラスタ選択指示工程により選択された部分集合に属する文書の文書特徴ベクトルを除去したのりとなるように修正するベクトル修正工程と、

を含み、

前記分類工程は、前記ベクトル修正工程により修正された文書特徴ベクトルに基づいて文書を分類することを特徴とする請求項31に記載の文書分類方法。

【請求項33】 前記ベクトル生成工程により生成された文書特徴ベクトル間の類似度を判断する際の文書表現空間を前記クラスタ選択指示工程により選択された部分集合から算出する特微量に基づいて修正する文書表現空間修正工程と

を含み、

前記分類工程は、前記文書表現空間修正工程により修正された文書表現空間を用いて、前記ベクトル生成手段工程により生成された文書特徴ベクトル間の類似度に基づいて文書を分類することを特徴とする請求項31に記載の文書分類方法。

【請求項34】 前記ベクトル生成工程により生成された文書特徴ベクトル間の類似度を判断する際の文書表現空間を前記クラスタ選択指示工程により選択された部分集合から算出する特徴量に基づいて修正する文書表現空間修正工程とを含み、

前記分類工程は、前記文書表現空間修正工程により修正された文書表現空間を用いて、前記ベクトル修正工程により修正された文書特徴ベクトル間の類似度に基づいて文書を分類することを特徴とする請求項32に記載の文書分類方法。

【請求項35】 前記分類工程により生成された文書の部分集合に所属する文書のすべてあるいは一部が選択された場合に選択されたことを示す選択情報を付与する選択情報付与工程を含み、

前記表示工程は、前記クラスタ特徴を表示するとともに、前記選択情報付与工程により付与された選択情報を表示することを特徴とする請求項31または33に記載の文書分類方法。

【請求項36】 前記分類体系生成工程は、前記選択指示工程により選択されたクラスタ特徴のほか、前記文書の部分集合の中から選択された文書の部分集合に所属する文書群の全部あるいは一部および／または操作者が作成した情報に基づいて分類体系の構成要素を生成することを特徴とする請求項31～35に記載の文書分類方法。

【請求項37】 文書の内容にしたがって文書群を分類する文書分類方法において、

文書データ群を入力し、入力された文書データ群の各文書に対して所定の基準に基づき文書の分割をおこない、一つの文書データから一つまたは複数の分割文書データを生成し、前記文書データと前記分割文書データとの対応を示す文書一分割文書対応マップを生成し、前記分割文書データを分類し、分割文書分類結果

情報を生成し、前記文書一分割文書対応マップと前記分割文書分類結果情報とを用いて前記文書データの分類結果情報を生成することを特徴とする文書分類方法。

【請求項38】 前記請求項24～37のいずれか一つに記載された方法をコンピュータに実行させるプログラムを記録したことを特徴とするコンピュータ読み取り可能な記録媒体。

【発明の詳細な説明】

【0001】

【発明の属する技術分野】

この発明は、入力された複数の文書データを所定の形式で表示または印刷するために出力する文書処理装置、文書処理方法およびその方法をコンピュータに実行させるプログラムを記録したコンピュータ読み取り可能な記録媒体に関する。また、この発明は、入力された複数の文書をその文書の内容に基づいて分類をおこなう、特に文書分類の際に算出される分類カテゴリ（体系）を精錬化する文書分類装置、文書分類方法およびその方法をコンピュータに実行させるプログラムを記録したコンピュータ読み取り可能な記録媒体に関する。

【0002】

【従来の技術】

近年、さまざまな文書分類装置や文書検索装置が開発されている。また、インターネット等のネットワーク技術の普及により国内外の大量の電子化文書へのアクセスが可能になり、それに比例して業務上電子的に蓄積される情報の量も飛躍的に拡大した。その中で収集した大量の文書情報を意味あるカテゴリ（体系）に分類する等の知的作業の必要性が高まっている。

【0003】

これらの大量の文書情報を意味的に分類するという作業の目的は、以下のようなものである。まず第1に、検索容易性の向上が考えられる。これは、膨大な文書群を分類名称（内容名）を手がかりに検索できるので検索が比較的容易になるというものである。

【0004】

第2に、情報群全体の把握が考えられる。これは、文書群全体がどのような内容（個々の分類）で構成されているかを把握する。しかし、大量の文書情報を操作者が手動で分類する場合、正確な分類をすることはできるが、分類に係る人的・時間的コストが膨大なものになるため、近年の文書の蓄積量の膨大さから、文書情報の自動分類装置が提案されるようになってきた。

【0005】

文書自動分類装置の従来技術としては、たとえば、特開平7-36897号公報に記載されているように、文書を、単語を特徴とする文書ベクトルとみなし、クラスタリング手法を用いてこれらの文書ベクトルを群分けし、群分けした文書ベクトルに基づいて文書の自動分類をおこなうものがある。

【0006】

また、「Projections for Efficient Document Clustering (著者名: Hinrich Schutze and Craig Silverstein, 学会名: ACM, 論文名: Proceedings of SIGIR, ページ: 74-81, 発行年: 1997)」においては、潜在的意味空間において文書分類を実施しているものがある。そのほかの方法としては、確率論的アプローチを用いる方法等が考えられる。

【0007】

また近年、インターネットなどの普及により、大量の文書群へのアクセスが可能になり、その結果、その文書群をさまざまな利用者の意図に基づいて、かつ、効率的に利用できるようにする必要性が高まっている。そのため、大量の文書群を意味のあるカテゴリに分類し、文書群の構造を把握するという知的作業がおこなわれ始めている。しかし、このような分類作業を人手によりおこなう場合、その人的および時間的なコストが膨大なものになるし、また、分類のための知識を分類者のみが有することになるため、分類担当者が代わると分類基準も変わってしまうことになる。

【0008】

そのため、文書群を人間が分類するような分類基準で自動的に分類しうる文書

分類装置が望まれており、文書分類装置としては、たとえば、特開平7-114572号公報に記載されているように、文書から自動的に単語の特徴ベクトルを抽出し、その特徴ベクトルをもとに文書分類することで、意味的な異なりを用いた自動分類を可能にするものがある。

【0009】

【発明が解決しようとする課題】

しかしながら、上記従来技術の文書分類装置は、本質的には単語で構成される多次元空間に布置した文書を統計的な分類をする方法であるため、分類結果は単語のいわゆる振る舞いという観点から統計的に求められたものにすぎず、分類の結果、算出される各クラスタ（分類された個々の文書の部分集合）が操作者（利用者）に理解不能な場合がある。

【0010】

また、どのような分類結果が最適かは、分類対象の文書集合の特徴や、利用者の作業の目的に依存するため、最適な分類結果について定義することが困難であるという問題点があった。特に、上記情報群全体の把握に関し、多様な操作者の意図により要求される分類も異なるため、一度の分類作業で、操作者の所望する結果を得ることが困難であるという問題点があった。

【0011】

このように、文書分類の結果は、多くのいわゆるノイズを含んだものであると解釈することができ、その一部についてのみが操作者にとって有益な場合が多いという問題点があった。

【0012】

また、これらの従来技術においては、文書の構成単位を考慮していないため、文書が一つまたは複数の段落記号やタイトルなどにより区切られた構造を持つ場合には、一つの文書の中に複数の話題や意味が含まれてしまい、その結果、利用者がその分類カテゴリを理解し難くなったり、また、ある特定の話題や特定の意味に限定されたカテゴリになったり、利用者の意図するカテゴリとは異なるカテゴリに分類されてしまうという問題が生じている。

【0013】

なお、特開平6-176064号公報に示された文脈依存自動分類装置には、文書の段落情報を考慮した文書自動分類をおこなうことにより分類精度を高めようとするものが開示されているか、本質的に上記の問題を解決するものではない。

【0014】

また、上記従来技術の文書分類装置や文書検索装置等の文書処理装置は、単に文書を分類する、あるいは文書を検索する機能を有するのみで、その結果を用いてさらなる分析をおこない、文書群に内在する隠れた情報の解析をおこなうことについては何ら考慮がされておらず、文書群に内在する隠れた情報の解析は別の解析装置を用いておこなわなければならないという問題点があった。

【0015】

また、情報分析をおこなう操作者が分類作業や検索作業をおこなうのは、これらの作業において、結果は目的なのではなく、単に情報分析作業の途中経過にすぎないからである。通常は、その後、さらに結果を把握しやすくするために、元の文書に含まれる情報を最大限に活用し、結果の並べ替えをおこなったり、集計・統計処理を施したり、結果をもとに表の形式にまとめたり、さらにはグラフ化したりというようなさまざまな処理を繰り返しおこない、意味ある情報分析結果を導き出す必要がある。

【0016】

また、数値データを対象とする情報の分析作業において、表計算ソフトウェアが用いられる場合があるが、表計算ソフトウェアは、元来、数値データの取扱いを意図して開発されたものであり、文字データ、特に文書の意味に係わるような分析作業においては十分な効果を発揮することはできなかった。

【0017】

この発明は、上述した従来例による問題点を解消するため、文書の意味に係わるような分析作業において、単に分類作業や検索作業などを固定された機能としておこない、その結果を出力するのではなく、情報分析作業全般にわたる支援をおこなうことができる文書処理装置、文書処理方法およびその方法をコンピュータに実行させるプログラムを記録したコンピュータ読み取り可能な記録媒体を提

供することを第1の目的とする。

【0018】

またこの発明は、上述した従来例による問題点を解消するため、任意の文書集合にどのような内容が含まれるかを漸次的に収集することができる文書分類装置、文書分類方法およびその方法をコンピュータに実行させるプログラムを記録したコンピュータ読み取り可能な記録媒体を提供することを第2の目的とする。

【0019】

またこの発明は、上述した従来例による問題点を解決するため、一つの文書の中に複数の話題や意味が含まれている場合に、ある特定の話題や意味に限定されたカテゴリに分類されたり、利用者の意図するカテゴリとは異なるカテゴリに分類されることがないことにより、利用者がその分類カテゴリをよく理解できる文書分類装置、文書分類方法およびその方法をコンピュータに実行させるプログラムを記録したコンピュータ読み取り可能な記録媒体を提供することを第3の目的とする。

【0020】

【課題を解決するための手段】

上述した課題を解決し、目的を達成するため、請求項1の発明に係る文書処理装置は、入力された複数の文書データを所定の形式で表示または印刷するために出力する文書処理装置において、入力された文書データを記憶する文書記憶手段と、前記文書記憶手段により記憶された文書データの全部または一部を選択する選択手段と、前記選択手段により選択された文書データの全部または一部から文字列の特徴に関するデータを抽出する特徴抽出手段と、前記特徴抽出手段により抽出された文字列の特徴に関するデータに基づいて前記文書データの全部または一部を加工処理する加工処理手段と、前記加工処理手段により加工処理された文書データの全部または一部を出力する出力手段と、を備えたことを特徴とする。

【0021】

この請求項1の発明によれば、文書の意味に係わるような分析作業において、単にその結果のみを出力するのではなく、情報分析作業全般にわたる支援をおこなうことができる。

【0022】

また、請求項2の発明に係る文書処理装置は、請求項1の発明において、前記出力手段が、前記加工処理手段により加工処理された文書データの全部または一部の内容に基づいて複数の項目値を設定する項目値設定手段と、前記項目値設定手段により設定された項目値ごとに前記文書データの全部または一部を集計する集計手段と、を備え、前記文書データの全部または一部を、項目値を少なくとも一つの軸とする表形式に展開して出力することを特徴とする。

【0023】

この請求項2の発明によれば、簡易な操作で加工処理の結果をクロス表として表すことができ、情報の内容の把握を容易におこなうことができることから、文書の意味に係わるような分析作業において、単にその結果のみを出力するのではなく、情報分析作業全般にわたる支援をおこなうことができる。

【0024】

また、請求項3の発明に係る文書処理装置は、請求項1または2の発明において、前記出力手段が、さらに、前記加工処理手段により加工処理された文書データの全部または一部を、前記加工処理手段により加工処理される前の文書データの全部または一部とともに出力することを特徴とする。

【0025】

この請求項3の発明によれば、加工処理すべき対象データとその他のデータが同時に表示され、それを確認することにより、加工処理の対象範囲の決定を正確かつ容易におこなうことができることから、文書の意味に係わるような分析作業において、単にその結果のみを出力するのではなく、情報分析作業全般にわたる支援をおこなうことができる。

【0026】

また、請求項4の発明に係る文書処理装置は、請求項1～3の発明において、前記文書記憶手段が、さらに、前記加工処理手段により加工処理された文書データの全部または一部を記憶することを特徴とする。

【0027】

この請求項4の発明によれば、以後、他のデータと同様に扱うことができるこ

とから、文書の意味に係わるような分析作業において、単にその結果のみを出力するのではなく、情報分析作業全般にわたる支援をおこなうことができる。

【0028】

また、請求項5の発明に係る文書処理装置は、請求項1～4の発明において、前記選択手段が、さらに、前記出力手段により出力された文書データの全部または一部を選択することを特徴とする。

【0029】

この請求項5の発明によれば、出力手段により出力された文書データの全部または一部をさらなる分析の対象とすることができます、多彩で高度な情報分析作業ができることから、文書の意味に係わるような分析作業において、単にその結果のみを出力するのではなく、情報分析作業全般にわたる支援をおこなうことができる。

【0030】

また、請求項6の発明に係る文書処理装置は、請求項1～5の発明において、前記文書記憶手段が、さらに、前記加工処理の内容に関するデータを記憶することを特徴とする。

【0031】

この請求項6の発明によれば、加工処理の内容に関するデータの紛失を防止し、当該データの管理が容易になるだけでなく、加工処理に用いた設定とそれによる処理結果を関連づけて把握することができることから、文書の意味に係わるような分析作業において、単にその結果のみを出力するのではなく、情報分析作業全般にわたる支援をおこなうことができる。

【0032】

また、請求項7の発明に係る文書分類装置は、文書の内容に基づいて文書の分類をおこなう文書分類装置において、文書データを入力する入力手段と、前記入力手段により入力された文書データを解析して言語解析情報を得る言語解析手段と、前記言語解析手段により得られた言語解析情報に基づいて前記文書データに対する文書特徴ベクトルを生成するベクトル生成手段と、前記ベクトル生成手段により生成された文書特徴ベクトル間の類似度に基づいて文書を分類し、文書の

部分集合を生成する分類手段と、前記分類手段により生成された文書の部分集合の特徴であるクラスタ特徴を算出するクラスタ特徴算出手段と、前記クラスタ特徴算出手段により算出されたクラスタ特徴を分類体系の構成要素として記憶する分類体系記憶手段と、を備えたことを特徴とする。

【0033】

この請求項7の発明によれば、クラスタを得ることができるとともに、クラスタ重心間の類似度等を用いて、クラスタの内容に基づくクラスタの構造化・体系化をおこなうことができる。

【0034】

また、請求項8の発明に係る文書分類装置は、文書の内容に基づいて文書の分類をおこなう文書分類装置において、文書データを入力する入力手段と、前記入力手段により入力された文書データを解析して言語解析情報を得る言語解析手段と、前記言語解析手段により得られた言語解析情報に基づいて前記文書データに対する文書特徴ベクトルを生成するベクトル生成手段と、前記ベクトル生成手段により生成された文書特徴ベクトル間の類似度に基づいて文書を分類し、文書の部分集合を生成する分類手段と、前記分類手段により生成された文書の部分集合の特徴であるクラスタ特徴を算出するクラスタ特徴算出手段と、前記クラスタ特徴算出手段により算出されたクラスタ特徴を表示する表示手段と、前記分類手段により生成された文書の部分集合の中から所望の部分集合を選択するクラスタ選択指示手段と、前記クラスタ選択指示手段により選択された文書の部分集合を分類体系の構成要素として記憶する分類体系記憶手段と、を備えたことを特徴とする。

【0035】

この請求項8の発明によれば、選択されたクラスタのみを用いており、より操作者の意図したものに近いクラスタの構造化・体系化をおこなうことができる。

【0036】

また、請求項9の発明に係る文書分類装置は、請求項8の発明において、前記ベクトル生成手段により生成された文書特徴ベクトルを記憶する文書特徴ベクトル記憶手段と、前記文書特徴ベクトル記憶手段により記憶された文書特徴ベクト

ルを、前記クラスタ選択指示手段により選択された部分集合に属する文書の文書特徴ベクトルを除去したのこりとなるように修正するベクトル修正手段と、を備え、前記分類手段が、前記ベクトル修正手段により修正された文書特徴ベクトルに基づいて文書を分類することを特徴とする。

【0037】

この請求項9の発明によれば、既知になったクラスタの影響を排除した新たなクラスタを生成することができる。

【0038】

また、請求項10の発明に係る文書分類装置は、請求項8の発明において、前記ベクトル生成手段により生成された文書特徴ベクトルを記憶する文書特徴ベクトル記憶手段と、前記文書特徴ベクトル記憶手段により記憶された文書特徴ベクトル間の類似度を判断する際の文書表現空間を前記クラスタ選択指示手段により選択された部分集合から算出する特徴量に基づいて修正する文書表現空間修正手段と、を備え、前記分類手段が、前記文書表現空間修正手段により修正された文書表現空間を用いて、前記ベクトル生成手段により生成された文書特徴ベクトル間の類似度に基づいて文書を分類することを特徴とする。

【0039】

この請求項10の発明によれば、前回の分類実行の結果、操作者に選択されたクラスタの形成特徴を次回の分類実行時に排除することができ、排除した状態で新たなクラスタを生成することができる。

【0040】

また、請求項11の発明に係る文書分類装置は、請求項9の発明において、前記ベクトル生成手段により生成された文書特徴ベクトルを記憶する文書特徴ベクトル記憶手段と、前記文書特徴ベクトル記憶手段により記憶された文書特徴ベクトル間の類似度を判断する際の文書表現空間を前記クラスタ選択指示手段により選択された部分集合から算出する特徴量に基づいて修正する文書表現空間修正手段と、を備え、前記分類手段が、前記文書表現空間修正手段により修正された文書表現空間を用いて、前記ベクトル修正手段により修正された文書特徴ベクトル間の類似度に基づいて文書を分類することを特徴とする。

【0041】

この請求項11の発明によれば、既知になったクラスタの影響を排除し、かつ、前回の分類実行の結果、操作者に選択されたクラスタの形成特徴を次回の分類実行時に排除することができ、排除した状態で新たなクラスタを生成することができる。

【0042】

また、請求項12の発明に係る文書分類装置は、請求項8または10の発明において、前記分類手段により生成された文書の部分集合に所属する文書のすべてあるいは一部が選択された場合に選択されたことを示す選択情報を付与する選択情報付与手段を備え、前記表示手段が、前記クラスタ特徴を表示するとともに、前記選択情報付与手段により付与された選択情報を表示することを特徴とする。

【0043】

この請求項12の発明によれば、多重に利用される文書の識別性および一度も選択されない文書の識別性を向上させることができる。

【0044】

また、請求項13の発明に係る文書分類装置は、請求項8～12の発明において、前記分類体系記憶手段が、前記選択指示手段により選択された文書の部分集合に属する全部あるいは一部の文書のほか、クラスタ特徴および／または操作者が作成した任意の情報を分類体系の構成要素として記憶することを特徴とする。

【0045】

この請求項13の発明によれば、クラスタの内容把握を容易にし、かつ、操作者独自の分類体系を簡易に生成できるので、分類体系の利用価値を向上させることができるものである。

【0046】

また、請求項14の発明に係る文書分類装置は、文書の内容にしたがって文書群を分類する文書分類装置において、文書データ群を入力する文書入力手段と、入力された文書データ群の各文書に対して所定の基準に基づき文書の分割をおこない、一つの文書データから一つまたは複数の分割文書データを生成する文書分割手段と、前記文書データと前記分割文書データとの対応を示す文書-分割文書

対応マップを生成する文書－分割文書対応マップ生成手段と、前記分割文書データを分類する分割文書分類手段と、前記分割文書分類手段による分類結果に基づいて分割文書分類結果情報を生成する分割文書分類結果生成手段と、前記文書－分割文書対応マップと前記分割文書分類結果情報を用いて前記文書データの分類結果情報を生成する文書分類結果生成手段と、を備えたことを特徴とする。

【0047】

この請求項14の発明によれば、一つの文書の中に複数の話題や意味が含まれている場合に、ある特定の話題や意味に限定されたカテゴリに分類されたり、利用者の意図するカテゴリとは異なるカテゴリに分類されたりすることがなく、したがって、利用者がその分類カテゴリをよく理解できる。また、分割前文書（所属文書）中の分割文書の位置なども示されるので、利用者は文書群中の読みたい部分を効率的に読むことができる。

【0048】

また、請求項15の発明に係る文書分類装置は、請求項14の発明において、前記文書データを保存する文書保存手段と、前記分割文書データを保存する分割文書保存手段と、前記文書－分割文書対応マップ生成手段により生成された文書－分割文書対応マップを保存する文書－分割文書対応マップ保存手段と、を備えたことを特徴とする。

【0049】

この請求項15の発明によれば、分割文書データおよび文書－分割文書対応マップを再生成することなしに、同一の文書データに対して、分類数、分類手法、または分類時の諸設定などパラメータの異なる分類結果を効率的に求めることができる。また、文書データを分類し、分類結果を生成するために必要なデータが保存されることにより、利用者が分類作業に対して時間的な自由度を持つことができるし、過去に行った文書分類の再分析を任意の時間間におこなうこともできる。

【0050】

また、請求項16の発明に係る文書分類装置は、請求項15の発明において、前記分割文書分類結果生成手段により生成された分割文書分類結果情報を保存す

る分割文書分類結果保存手段を備えたことを特徴とする。

【0051】

この請求項16の発明によれば、請求項15の発明の効果に加え、一度分類を実行すれば、その分類結果をテキスト表現や表表現やグラフ表現などさまざまな形式で表現することができる。また、分割文書分類結果情報が保存されることにより、分類の実行作業および分類結果の分析作業において、利用者が時間的な自由度を持つことができるし、過去に行った文書分類結果の再分析をさまざまな表現形式で任意の時間におこなうこともできる。

【0052】

また、請求項17の発明に係る文書分類装置は、請求項14～16の発明において、前記文書分割手段により生成される複数の分割文書データには分割前の文書データそのものを含むことを特徴とする。

【0053】

この請求項17の発明によれば、利用者は、分割されている文書データを分類することで得られる詳細な文書データの分類構造だけでなく、分割前の文書データ自体を分類した結果として得られる概略的でマクロな分類構造の融合した分類構造を得ることができる。

【0054】

また、請求項18の発明に係る文書分類装置は、請求項14～17の発明において、前記文書分割手段が、文書データの構造情報を基に文書データを分割する構成にしたことを特徴とする。

【0055】

この請求項18の発明によれば、異なった話題の分割等を適切におこなうことができ、したがって、文書データの詳細な分類構造がわかる文書分類を適切におこなうことができる。

【0056】

また、請求項19の発明に係る文書分類装置は、請求項14～17の発明において、前記文書データに含まれる要素を抽出する文書要素抽出手段と、前記文書要素抽出手段により抽出された要素に付随する要素付隨情報を抽出する要素付隨

情報抽出手段と、を備え、前記文書分割手段が、前記文書要素抽出手段により抽出された要素、または前記要素と前記要素付隨情報抽出手段により抽出された要素付隨情報とを用いて前記文書データを分割する構成にしたことを特徴とする。

【0057】

この請求項19の発明によれば、文書データの詳細な分類構造がわかる文書分類を適切におこなうことができる。

【0058】

また、請求項20の発明に係る文書分類装置は、請求項14～17の発明において、前記文書分割手段が、指示された指定範囲にしたがって文書データの分割をおこなう構成にしたことを特徴とする。

【0059】

この請求項20の発明によれば、利用者の意図に合い、かつ文書データの詳細な分類構造がわかる文書分類をおこなうことができる。

【0060】

また、請求項21の発明に係る文書分類装置は、請求項14～17において、前記文書分割手段が、文書データ中の文字数、文数、または文字数と文数の両方を基に文書データを分割する構成にしたことを特徴とする。

【0061】

この請求項21の発明によれば、話題の異なった内容などが異なった文書として分類される可能性が高くなり、したがって、この発明でも文書データの詳細な分類構造がわかる文書分類をおこなうことができる。

【0062】

また、請求項22の発明に係る文書分類装置は、請求項14～21の発明において、前記文書分類結果生成手段が、文書データを示す情報および前記文書データに付随する代表的情報を、分類結果情報として抽出して提示する構成にしたことを特徴とする。

【0063】

この請求項22の発明によれば、利用者は文書データの詳細な分類構造の概要や全体的な構造を容易に把握することができる。

【0064】

また、請求項23の発明に係る文書分類装置は、請求項22の発明において、前記文書分類結果生成手段が、分割文書データを示す情報および前記分割文書データに付随する代表的情報を、分類結果情報として、抽出して提示する構成にしたことを特徴とする。

【0065】

この請求項23の発明によれば、利用者は文書データの詳細な分類構造の概要や全体的な構造とともにどの分割文書が起因して当該カテゴリに分類されたかというようなことも容易にわかる。

【0066】

また、請求項24の発明に係る文書処理方法は、入力された複数の文書データを所定の形式で表示または印刷するために出力する文書処理方法において、入力された文書データを記憶する文書記憶工程と、前記文書記憶工程により記憶された文書データの全部または一部を選択する選択工程と、前記選択工程により選択された文書データの全部または一部から文字列の特徴に関するデータを抽出する特徴抽出工程と、前記特徴抽出工程により抽出された文字列の特徴に関するデータに基づいて前記文書データの全部または一部を加工処理する加工処理工程と、前記加工処理工程により加工処理された文書データの全部または一部を出力する出力工程と、を含んだことを特徴とする。

【0067】

この請求項24の発明によれば、文書の意味に係わるような分析作業において、単にその結果のみを出力するのではなく、情報分析作業全般にわたる支援をおこなうことができる。

【0068】

また、請求項25の発明に係る文書処理方法は、請求項24の発明において、前記出力工程が、前記加工処理工程により加工処理された文書データの全部または一部の内容に基づいて複数の項目値を設定する項目値設定工程と、前記項目値設定工程により設定された項目値ごとに前記文書データの全部または一部を集計する集計工程と、を含み、前記文書データの全部または一部を、項目値を少なく

とも一つの軸とする表形式に展開して出力することを特徴とする。

【0069】

この請求項25の発明によれば、簡易な操作で加工処理の結果をクロス表として表すことができ、情報の内容の把握を容易におこなうことができることから、文書の意味に係わるような分析作業において、単にその結果のみを出力するのではなく、情報分析作業全般にわたる支援をおこなうことができる。

【0070】

また、請求項26の発明に係る文書処理方法は、請求項24または25の発明において、前記出力工程が、さらに、前記加工処理工程により加工処理された文書データの全部または一部を、前記加工処理工程により加工処理される前の文書データの全部または一部とともに出力することを特徴とする。

【0071】

この請求項26の発明によれば、加工処理すべき対象データとその他のデータが同時に表示され、それを確認することにより、加工処理の対象範囲の決定を正確かつ容易におこなうことができることから、文書の意味に係わるような分析作業において、単にその結果のみを出力するのではなく、情報分析作業全般にわたる支援をおこなうことができる。

【0072】

また、請求項27発明に係る文書処理方法は、請求項24～26の発明において、前記文書記憶工程が、さらに、前記加工処理工程により加工処理された文書データの全部または一部を記憶することを特徴とする。

【0073】

この請求項27の発明によれば、以後、他のデータと同様に扱うことができることから、文書の意味に係わるような分析作業において、単にその結果のみを出力するのではなく、情報分析作業全般にわたる支援をおこなうことができる。

【0074】

また、請求項28の発明に係る文書処理方法は、請求項24～27の発明において、前記選択工程が、さらに、前記出力工程により出力された文書データの全部または一部を選択することを特徴とする。

【0075】

この請求項28の発明によれば、出力手段により出力された文書データの全部または一部をさらなる分析の対象とすることができます、多彩で高度な情報分析作業ができることから、文書の意味に係わるような分析作業において、単にその結果のみを出力するのではなく、情報分析作業全般にわたる支援をおこなうことができる。

【0076】

また、請求項29の発明に係る文書処理方法は、請求項24～28の発明において、前記文書記憶工程が、さらに、前記加工処理の内容に関するデータを記憶することを特徴とする。

【0077】

この請求項29の発明によれば、加工処理の内容に関するデータの紛失を防止し、当該データの管理が容易になるだけでなく、加工処理に用いた設定とそれによる処理結果を関連づけて把握することができることから、文書の意味に係わるような分析作業において、単にその結果のみを出力するのではなく、情報分析作業全般にわたる支援をおこなうことができる。

【0078】

また、請求項30の発明に係る文書分類方法は、文書の内容に基づいて文書の分類をおこなう文書分類方法において、文書データを入力する入力工程と、前記入力工程により入力された文書データを解析して言語解析情報を得る言語解析工程と、前記言語解析工程により得られた言語解析情報に基づいて前記文書データに対する文書特徴ベクトルを生成するベクトル生成工程と、前記ベクトル生成工程により生成された文書特徴ベクトル間の類似度に基づいて文書を分類し、文書の部分集合を生成する分類工程と、前記分類工程により生成された文書の部分集合の特徴であるクラスタ特徴を算出するクラスタ特徴算出工程と、前記クラスタ特徴算出工程により算出されたクラスタ特徴に基づいて分類体系の構成要素を生成する分類体系生成工程と、を含んだことを特徴とする。

【0079】

この請求項30の発明によれば、クラスタを得ることができるとともに、クラ

クラスタ間の類似度等を用いて、クラスタの内容に基づくクラスタの構造化・体系化をおこなうことができる。

【0080】

また、請求項31の発明に係る文書分類方法は、文書の内容に基づいて文書の分類をおこなう文書分類方法において、文書データを入力する入力工程と、前記入力工程により入力された文書データを解析して言語解析情報を得る言語解析工程と、前記言語解析工程により得られた言語解析情報に基づいて前記文書データに対する文書特徴ベクトルを生成するベクトル生成工程と、前記ベクトル生成工程により生成された文書特徴ベクトル間の類似度に基づいて文書を分類し、文書の部分集合を生成する分類工程と、前記分類工程により生成された文書の部分集合の特徴であるクラスタ特徴を算出するクラスタ特徴算出工程と、前記クラスタ特徴算出工程により算出されたクラスタ特徴を表示する表示工程と、前記分類工程により生成された文書の部分集合の中から所望の部分集合を選択するクラスタ選択指示工程と、前記クラスタ選択指示工程により選択されたクラスタ特徴に基づいて分類体系の構成要素を生成する分類体系生成工程と、を含んだことを特徴とする。

【0081】

この請求項31の発明によれば、選択されたクラスタのみを用いており、より操作者の意図したものに近いクラスタの構造化・体系化をおこなうことができる。

【0082】

また、請求項32の発明に係る文書分類方法は、請求項31の発明において、前記ベクトル生成工程により生成された文書特徴ベクトルを、前記クラスタ選択指示工程により選択された部分集合に属する文書の文書特徴ベクトルを除去したのこりとなるように修正するベクトル修正工程と、を含み、前記分類工程が、前記ベクトル修正工程により修正された文書特徴ベクトルに基づいて文書を分類することを特徴とする。

【0083】

この請求項32の発明によれば、既知になったクラスタの影響を排除した新た

なクラスタを生成することができる。

【0084】

また、請求項33の発明に係る文書分類方法は、請求項31の発明において、前記ベクトル生成工程により生成された文書特徴ベクトル間の類似度を判断する際の文書表現空間を前記クラスタ選択指示工程により選択された部分集合から算出する特微量に基づいて修正する文書表現空間修正工程と、を含み、前記分類工程が、前記文書表現空間修正工程により修正された文書表現空間を用いて、前記ベクトル生成手段工程により生成された文書特徴ベクトル間の類似度に基づいて文書を分類することを特徴とする。

【0085】

この請求項33の発明によれば、前回の分類実行の結果、操作者に選択されたクラスタの形成特徴を次回の分類実行時に排除することができ、排除した状態で新たなクラスタを生成することができる。

【0086】

また、請求項34の発明に係る文書分類方法は、請求項32の発明において、前記ベクトル生成工程により生成された文書特徴ベクトル間の類似度を判断する際の文書表現空間を前記クラスタ選択指示工程により選択された部分集合から算出する特微量に基づいて修正する文書表現空間修正工程と、を含み、前記分類工程が、前記文書表現空間修正工程により修正された文書表現空間を用いて、前記ベクトル修正工程により修正された文書特徴ベクトル間の類似度に基づいて文書を分類することを特徴とする。

【0087】

この請求項34の発明によれば、既知になったクラスタの影響を排除し、かつ、前回の分類実行の結果、操作者に選択されたクラスタの形成特徴を次回の分類実行時に排除することができ、排除した状態で新たなクラスタを生成することができる。

【0088】

また、請求項35の発明に係る文書分類方法は、請求項31または33の発明において、前記分類工程により生成された文書の部分集合に所属する文書のすべ

てあるいは一部が選択された場合に選択されたことを示す選択情報を付与する選択情報付与工程を含み、前記表示工程が、前記クラスタ特徴を表示するとともに、前記選択情報付与工程により付与された選択情報を表示することを特徴とする。

【0089】

この請求項35の発明によれば、多重に利用される文書の識別性および一度も選択されない文書の識別性を向上させることができる。

【0090】

また、請求項36の発明に係る文書分類方法は、請求項31～35の発明において、前記分類体系生成工程が、前記選択指示工程により選択されたクラスタ特徴のほか、前記文書の部分集合の中から選択された文書の部分集合に所属する文書群の全部あるいは一部および／または操作者が作成した情報に基づいて分類体系の構成要素を生成することを特徴とする。

【0091】

この請求項36の発明によれば、クラスタの内容把握を容易にし、かつ、操作者独自の分類体系を簡易に生成できるので、分類体系の利用価値を向上させることができることとなる。

【0092】

また、請求項37の発明に係る文書分類方法は、文書の内容にしたがって文書群を分類する文書分類方法において、文書データ群を入力し、入力された文書データ群の各文書に対して所定の基準に基づき文書の分割をおこない、一つの文書データから一つまたは複数の分割文書データを生成し、前記文書データと前記分割文書データとの対応を示す文書－分割文書対応マップを生成し、前記分割文書データを分類し、分割文書分類結果情報を生成し、前記文書－分割文書対応マップと前記分割文書分類結果情報を用いて前記文書データの分類結果情報を生成することを特徴とする。

【0093】

この請求項37の発明によれば、一つの文書の中に複数の話題や意味が含まれている場合に、ある特定の話題や意味に限定されたカテゴリに分類されたり、利

用者の意図するカテゴリとは異なるカテゴリに分類されたりすることなく、したがって、利用者がその分類カテゴリをよく理解できる。また、分割前文書（所属文書）中の分割文書の位置なども示されるので、利用者は文書群中の読みたい部分を効率的に読むことができる。

【0094】

また、請求項38の発明に係る記憶媒体は、請求項24～37に記載された方法をコンピュータに実行させるプログラムを記録したことで、そのプログラムを機械読み取り可能となり、これによって、請求項24～37の動作をコンピュータによって実現することが可能である。

【0095】

【発明の実施の形態】

以下に添付図面を参照して、この発明に係る文書処理装置、文書処理方法およびその方法をコンピュータに実行させるプログラムを記録したコンピュータ読み取り可能な記録媒体の好適な実施の形態を詳細に説明する。

【0096】

【実施の形態1】

まず、この発明の実施の形態1による文書処理装置を構成する情報処理システム全体のハードウェア構成を説明する。図1は、実施の形態1による文書処理装置を構成する情報処理システム全体のハードウェア構成を示す説明図である。

【0097】

図1において、実施の形態1による文書処理装置を構成する情報処理システムは、サーバー／クライアント方式で構成されている。すなわち、サーバー101と複数のクライアント102がネットワーク103によって接続されている。クライアント102は、分類データ等の加工データの生成、サーバー101への指示、分類結果等の加工処理結果の表示などを起こす。一方、クライアント102からの指示にしたがって、サーバー101は文書（テキスト）分類等の加工処理を膨大な数値演算によりおこない、その処理の結果をクライアント102へ送る。

【0098】

分類処理の場合、より具体的には、サーバー101においては、テキスト分類処理（前処理、クラスタリング処理）がおこなわれ、クライアント102においては、分類データ生成、処理実行指示、テキスト分類結果表示等がおこなわれる。サーバー101における処理は、上述のように、「前処理」と「分類処理」の二つに分かれており、その処理はデータによっては非常に負荷が大きくなる。したがって、サーバー101は「前処理」と「分類処理」がそれぞれ一つずつしか処理をおこなわないようにマネージャプロセスが処理受付リストを作成して管理する。

【0099】

また、サーバー101とクライアント102との間のデータのやりとりはファイル共有という方法を用いる。すなわち、分類処理等の加工処理に用いるファイルをサーバー101上の共有フォルダに作成することにより両者はデータのやりとりをおこなう。したがって、クライアント102からはサーバー101の共有フォルダをネットワーク共有して利用することが可能である。

【0100】

つぎに、サーバー101およびクライアント102のハードウェア構成について説明する。図2は、実施の形態1による文書処理装置を構成する情報処理システムにおけるサーバー101のハードウェア構成を示す説明図である。サーバー101は、たとえばワークステーション(WS)等が用いられる。

【0101】

図2において、201はサーバー101全体を制御するCPUを、202はブートプログラム等を記憶したROMを、203はCPU201のワークエリアとして使用されるRAM203を、204は通信回線205を介してネットワーク103に接続され、そのネットワーク103と内部のインターフェイスを司るインターフェイス(I/F)を、206はデータを記憶するディスク装置を示している。200は上記各部を結合させるためのバスを示している。

【0102】

そのほか、文書情報、画像情報、機能情報等を表示するディスプレイ208や、データを入力するためのキーボード209およびマウス210等が同様に接続

されていてもよい。さらに、ディスク装置206には、クライアント102との間のデータのやりとりをするための共有フォルダ207が設けられている。

【0103】

また、図3は、実施の形態1による文書処理装置を構成する情報処理システムにおけるクライアント102のハードウェア構成を示す説明図である。クライアント102は、たとえばパーソナルコンピュータ（PC）等が用いられる。

【0104】

図3において、301はシステム全体を制御するCPUを、302はブートプログラム等を記憶したROMを、303はCPU301のワークエリアとして使用されるRAMを、304はCPU301の制御にしたがってHD（ハードディスク）305に対するデータのリード／ライトを制御するHDD（ハードディスクドライブ）を、305はHDD304の制御で書き込まれたデータを記憶するHDを、306はCPU301の制御にしたがってFD（フロッピーディスク）307に対するデータのリード／ライトを制御するFDD（フロッピーディスクドライブ）を、307はFDD306の制御で書き込まれたデータを記憶する着脱自在のFDを、308はドキュメント、画像、機能情報等を表示するディスプレイをそれぞれ示している。

【0105】

また、309は通信回線310を介してネットワーク103に接続され、そのネットワーク103と内部のインターフェイスを司るインターフェイス（I/F）を、311は文字、数値、各種指示等の入力のためのキーを備えたキーボードを、312はカーソルの移動や範囲選択、あるいは表示画面に表示されたアイコンやボタンの押下やウインドウの移動やサイズの変更等をおこなうマウスを、313はOCR（Optical Character Reader）機能を備えた画像を光学的に読み取るスキヤナを、314は分類結果を含むデータの内容等を印刷するプリンタを、315は上記各部を結合するためのバスをそれぞれ示している。また、HD305にはワープロソフト等のアプリケーションソフト316が記憶されている。

【0106】

つぎに、実施の形態1による文書処理装置の機能的構成について説明する。図4は、実施の形態1による文書処理装置の構成を機能的に示すブロック図である。図4において、文書処理装置は、入力部401と、文書記憶部402と、選択部403と、特徴抽出部404と、加工処理部405と、出力部406を含む構成である。

【0107】

入力部401、文書記憶部402、選択部403、特徴抽出部404、加工処理部405、出力部406は、ROM202または302、RAM203または303、あるいはディスク装置306またはハードディスク316等の記録媒体に記録されたプログラムに記載された命令にしたがってCPU201または301等が命令処理を実行することにより、各部の機能を実現する。

【0108】

入力部401は、文書データを入力するものであり、たとえば、キーボード209または311、スキャナ313、OCR機能を備えたスキャナ313、またはネットワーク103を経由して文書や文書群を得ることができるI/F204または309等である。また、入力部401は、上記以外に、文書データを取得することができるものであれば、それらのすべてを含む。たとえば、文書データがデータベース化されている場合に、そのデータベースが記録された媒体を実施の形態1の文書処理装置に組み入れた場合も文書データの入力とする。

【0109】

ここで、文書とは、自然言語で記述された一つ以上の文の集まりであり、文字、文字列、数値等から構成されており、それらの意味のあるまとまりを一つの文書とする。また、複数の文書の集まりを、文書群とする。

【0110】

文書は一つあるいは複数の項目から構成されている。項目は、項目名と、項目値から構成されている。項目名は項目の内容を示すラベルであり、文書に含まれていても含まれていなくてもよい。項目値は項目の実際の内容である。図5は、実施の形態1による文書処理装置の項目名と項目値の関係を示す説明図である。たとえば、一つの特許公報は一つの文書であり、特許公報を項目名と項目値によ

って表現すると、図5のようになる。

【0111】

入力部401によって取得された文書あるいは文書群は、それぞれの文書に一意な文書IDが付与され、文書記憶部402により記憶される。図6は、実施の形態1による文書処理装置の文書記憶部402に記憶された文書のデータ構造を示す説明図である。各項目名あるいは項目値は、文書記憶部402のセル、すなわち一つの記憶単位に収納される。

【0112】

図6においては、一つのセルは3つの記憶領域から構成されており、第1番目の記憶領域601にはつぎのセルの文書記憶部402上の位置（番地）が記憶されている。第2番目の記憶領域602には、セルの属性値が記憶されている。

【0113】

セルの属性値としては、たとえば、「0」が「空」、「1」が「数値」、「2」が文字列・・・というように設定することができる。第3番目の記憶領域603には、セルの実際の内容、すなわち、項目名あるいは項目値等が格納される領域の先頭位置が記憶されている。

【0114】

セルの順序の並び替えや、セルの追加・削除は、第1番目の記憶領域601に記憶されたつぎのセルの位置を変更することにより、容易に実現することができる。また、セルの実際の内容は、セルの構造とは異なる別の領域に記憶されているので、たとえば、項目を変更した結果、あらかじめ確保された領域では收まり切れなくなった場合には、セルの構造自体には影響なく、別途大きな領域を確保してそこに項目値を記憶し、第3番目に記憶された記憶領域603の先頭位置を変更するだけよい。

【0115】

図7は、実施の形態1による文書処理装置の文書記憶部402に記憶された文書の別のデータ構造を示す説明図である。図7において、一つのセルは二つの記憶領域を使用している。第1番目の記憶領域701には、セル属性値が記憶されている。第2番目の記憶領域702には、セルの実際の内容、すなわち項目名あ

るいは項目値などが格納される領域の先頭位置が記憶されている。

【0116】

つぎのセルは、文書記憶部402上でとなり合うつぎの記憶領域に記憶されている。このデータ構造では、セルの順序の並び替え、セルの追加・削除が発生した場合には、記憶内容の移動操作が必要となる。

【0117】

文書記憶部402は、通常高速に情報を扱える半導体メモリで構成されるが、磁気ディスクあるいは光ディスク等で構成される補助記憶装置を含んでいてよい。

【0118】

文書記憶部402に記憶された文書あるいは文書群は、出力部406により表示される。実施の形態1においては、出力部406は、CRTディスプレイ、液晶ディスプレイ等から構成される。出力部406は、文書記憶部402に記憶された文書あるいは文書群の内容をセルと順次たどって読み出し、表の形式で表示または印刷する。

【0119】

また、出力部406は、表の形式で表示または印刷されたデータに基づいてグラフを描画するグラフ描画部407を含んでいてよい。グラフ描画部407は、文書記憶部402に記憶された文書あるいは文書群の項目値に対して利用者が設定した領域の内容を読み出し、利用者の指示により棒グラフ、円グラフ、折れ線グラフ等のグラフを描画し、表示または印刷する。

【0120】

出力部406は、入力部401による操作に関する表示、たとえば、操作メニューやマウスポインタ、カーソルの表示等もおこなう。また、処理結果を印刷するためのプリンタ等の印刷装置を含んでいてよい。

【0121】

選択部403は、入力部401による操作者の指示により、出力部406の表示上で選択された領域のデータを文書記憶部402から読み出し、特徴抽出部404へ送る。選択部403の選択方法について、図8～図10を用いて説明する

【0122】

図8～図10は、実施の形態1による文書処理装置の出力部406による画面表示の例、具体的には、自動車の故障状況の内容が表示された画面表示の例を示す説明図である。図8において、画面表示には、文書ID番号を示す「番号」欄801、故障情報を受け付けた日付を示す「受付日」欄802、故障情報を受け付けた営業所を示す「営業所」欄803、故障情報の対象となった自動車の車種を示す「車種」欄804、故障情報対象となった自動車の年式を示す「年式」欄805、故障状況の内容を示す「内容」欄806が表示される。

【0123】

図9において、選択領域901は、矩形で囲まれ、表示色が変更されている部分であり、図10においても同様に、選択領域1001は、矩形で囲まれ、表示色が変更されている部分である。

【0124】

選択部403が選択する領域としては、図9に示すように、画面上の列の一部であってもよいし、また、図10に示すように項目名を選択した場合はその項目名に属する項目値全部が選択されるようにしてもよい。なお、実施の形態1では、文字列の属性を持つ領域のみ選択可能とする。

【0125】

つぎに、特徴抽出部404によりおこなわれる抽出処理の内容について説明する。選択部403により選択された項目値は、特徴抽出部404によりその項目値の特徴が抽出される。図11は、実施の形態1による文書処理装置の特徴抽出部404によりおこなわれる抽出処理の内容の一覧を示す説明図である。

【0126】

図11において、抽出処理には、対象とする文字列に含まれる単語、その単語の単語数、単語の文字数、単語のそれぞれの出現回数. . . 等がある。これらの抽出処理は、規則音声合成装置や自動翻訳装置等の一般的に用いられている形態素解析技術あるいは構文解析技術等の自然言語処理技術を用いて実現する。

【0127】

つぎに、加工処理部405によりおこなわれる加工処理の内容について説明する。特徴抽出部404により抽出処理された特徴量に対して、加工処理部405により加工処理が施される。図12は、実施の形態1による文書処理装置の加工処理部405によりおこなわれる加工処理の内容の一覧を示す説明図である。

【0128】

加工処理には、同一の特徴量ごと分類する「分類処理」、所定の特徴量を検索する「検索処理」、特徴量の内容ごとに並べ替えをおこなう「並べ替え処理」、特徴量の代表値を抽出する「代表値抽出処理」、特徴量のうちの最大値を抽出する「最大値抽出処理」、特徴量のうち最小値を抽出する「最小値抽出処理」、特徴量を算術する「算術処理」等がある。

【0129】

特徴抽出部404によりおこなわれる特徴量の抽出処理の内容と、加工処理部405によりおこなわれる抽出された特徴量の加工処理の内容の組み合わせは、おのおの操作者が選択できるようにすることができる。また、効果の高い組み合わせをあらかじめ設定して、その設定された組み合わせを操作者に提供するようにもよい。

【0130】

加工処理部405により加工処理された処理結果は、加工処理部405内の加工処理結果保持部408に保持される。加工処理結果保持部408に保持された加工処理結果は、出力部406により出力される。出力部406は、加工処理結果保持部408から内容を読み出し、画像表示や印刷出力をおこなう。

【0131】

ここで、特徴抽出部404により抽出される特徴（量）として、項目値に含まれる単語それぞれの出現回数を選択し、加工処理部405によりおこなわれる加工処理として、分類処理を選択した場合について説明する。

【0132】

一般的に、二つの文書があり、それら二つの文書を構成する単語の出現頻度が等しい場合、それら二つの文書の意味は似通っていると考えることができる。すなわち、ある文書での単語の出現回数は、その文書の意味に関係の深い特徴量で

あると考えることができる。したがって、単語の出現回数を特徴量として、複数の文書を分類した場合、それぞれの分類カテゴリには意味の近い文書が所属すると考えることができる。

【0133】

選択部403により選択された一つあるいは複数の項目値は、特徴抽出部404に含まれる解析部409によって項目値ごとに形態素解析等の自然言語解析をおこない、単語に分割される。また、それぞれの単語には、その単語の品詞情報も付与される。出現した単語のうち、名詞であるものに対して一意な単語IDを付与し、一つの項目値および選択部403により選択されたすべての項目値に対する単語IDごとの出現回数を計数する。

【0134】

特徴抽出部404に含まれる特徴ベクトル生成部410は、計数された出現回数に基づいて個々の項目値の特徴（量）を示す項目値特徴ベクトルを生成する。たとえば、選択部403により選択された項目値が、

「騒音が大きい」

「塗装が変色する」

「オーバーヒートが起こる」

「塗装がはげる」

「バッテリーが上がる」

「排気が黒い」

であった場合、各項目の特徴ベクトルは、図13に示すようになる。また、図14には、単語とその単語IDごとの出現回数を示す。

【0135】

すなわち、

「騒音が大きい」 : {1, 1, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0}

「塗装が変色する」 : {0, 1, 0, 1, 1, 1, 0, 0, 0, 0, 0, 0, 0}

「オーバーヒートが起こる」 : {0, 1, 0, 0, 0, 0, 1, 1, 0, 0, 0, 0, 0}

, 0, 0 }

「塗装がはげる」 : {0, 1, 0, 1, 0, 0, 0, 0, 1, 0, 0}

, 0, 0 }

「バッテリーが上がる」 : {0, 1, 0, 0, 0, 0, 0, 0, 0, 1, 1}

, 0, 0 }

「排気が黒い」 : {0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0}

, 1, 1 }

という特徴ベクトルが得られる。

【0136】

この項目値の特徴ベクトルは、特徴抽出部404からの出力として加工処理部405へ送られる。加工処理部405においては、項目値の特徴ベクトルを用いて、分類処理をおこなう。分類処理は、まず、おのおののベクトル間の距離を計算することにより求める。距離の尺度には、たとえば内積を用いることができる。

【0137】

距離を計算した後、距離の近いものをまとめる処理をおこなう。この処理には、たとえばベクトルの集合をその距離に応じてK個のベクトルの集合に分類するK-means法を用いることができる。それぞれのベクトルの分類が完了したら、ベクトルに対応する項目値に対してどの分類に属するかの番号、すなわち、クラスタ番号と、項目値に対応する文書IDを付与し、加工処理部405の出力とし、出力部406により出力表示をおこなう。

【0138】

図15は、クラスタ番号1501を表示した画面の表示例を示す。クラスタ番号が同一番号である文書、たとえば、番号「1」および「6」の文書のクラスタ番号が「5」であり、両者が同一の分類に属することを示している。

【0139】

つぎに、請求項2の発明においてクロス表を出力する場合について説明する。入力部401により、分析対等とする文書群を読み込んだ後、操作者は分類処理をおこなう対象となる項目名、クロス表の横軸あるいは縦軸となる項目名、いく

つに分類をおこなうかの分類数を指示する。

【0140】

図16はクロス表作成のための指示画面である。図16において、指示画面1600は、処理対象項目名入力欄1601と、軸となる項目名入力欄1602と、縦軸指定ボタン1603と、横軸指定ボタン1604と、分類数入力欄1605とから構成される。

【0141】

処理対象項目名入力欄1601には、処理対象となる項目名を入力する。キーボード209等から入力するあるいは処理対象となる項目候補を表示させその中からマウス210等により選択することにより項目名を入力することができる。また、軸となる項目名入力欄1602には、軸となる項目名を入力する。入力の方法は、処理対象項目名入力欄1601への入力の方法と同様である。

【0142】

縦軸指定ボタン1603および横軸指定ボタン1604は、軸となる項目を縦軸に表示させるか横軸に表示させるかを指定するためのボタンである。また、分類数入力欄1605には、いくつに分類するかその分類数を入力する。入力の方法としては、キーボード209等から数字を入力するあるいは分類数候補を表示させその中からマウス210等により選択することにより分類数を入力するようにもよい。

【0143】

図16においては、処理対象項目名入力欄1601には「内容」が、軸となる項目名入力欄1602には「車種」が、また、横軸指定ボタン1604がチェックされ、分類数入力欄1605には「50」が入力され、これにより、文書群の中の「内容」に基づいて、「50(個)」に分類され、クロス表の横軸に「車種」を表示するという指示がなされていることがわかる。

【0144】

クロス表作成の指示がおこなわれることにより、分類処理が実行され、その結果がクロス表で表示される。図17および図18は、分類処理の結果が表示されたクロス表を示す図である。図17において、クロス表1700は、縦軸に分類

を示す「クラスタ1」、「クラスタ2」...が表示され、横軸に車種を示す「ABC1600」、「ABC1800」...が表示される。

【0145】

表の縦軸、すなわち各行は、分類処理により生成されたクラスタに対応する。各行の第1欄には、分類処理終了時には既定値としてクラスタ番号を示す文字列が入っている。表の横軸、すなわち各欄には、文書群の項目「車種」に含まれる文字列が重複することなく表示される。行「クラスタ1」の各セルには、クラスタ1に分類された文書のうち、項目「車種」の値がその欄の車種と一致するものの数が表示される。

【0146】

ここで、数を表示する代わりに、セルの色の濃淡や、セルを塗りつぶす面積により数の大きさを表現するようにしてもよい。また、表の最右欄および最下欄には、該当する行、欄の合計が表示される。

【0147】

図18において、クロス表1700のあるセルにマウスポインタ1800を移動させ、マウス210のマウスボタンを押下する、あるいはキーボード209のカーソルキー操作によりカーソルを移動させ、特定キーを押下すると、そのセルの近傍に内容表示画面1801が表示されることにより、該当する文書の項目「内容」が表示される。

【0148】

内容表示画面1801には、セル内のデータ数、表示項目、セル情報、および各データにおける表示項目の内容が表示される。マウスポインタ1800により指定されたセルにおいては、データ数：「4」、表示項目：「内容」、セル情報：「ABC2000-クラスタ1」、表示項目の内容として「内容」の4つの内容である「排気が黒い、排気が黒い、...」が表示される。これにより、マウスポインタを所望のセルに移動させてマウスボタンを押下するという簡易な操作により、セルの内容を認識することができる。

【0149】

また、内容表示画面1801に表示される項目は、設定操作により変更するこ

とが可能であり、すべての項目を表示させることもでき、また、項目を選択して表示させることもできる。

【0150】

各行の第1欄には、分類処理終了時には既定値としてクラスタ番号を示す文字列が入っているが、操作者により、この欄の書き換えをすることができる。たとえば、上記の操作によってセルの内容を確認した後、「クラスタ1」を「排気の問題」と書き換えることができる。これにより、情報内容の把握がより容易になる。

【0151】

また、分類終了時に既定値としてクラスタ番号を示す文字列を入れるのではなく、そのクラスタの特徴を示す文字列を抽出し、セルに入れることも可能である。たとえば、クラスタ1に含まれる文書の項目「内容」から、もっとも頻度が高く出現する文や単語を抽出することにより実現する。

【0152】

図18においては、クラスタ1には「排気が黒い」あるいは「排気」等の単語が入れられる。このように、操作者は簡易な操作により文書全体の分布状態を把握するだけでなく、必要に応じて個々の文書の内容をも詳細に知ることができる。

【0153】

つぎに、クロス表を作成するための出力部406の詳細な構成の内容について説明する。図19は、実施の形態1による文書処理装置の出力部406の詳細な構成を示すブロック図である。出力部406は、グラフ描画部407のほかに、項目値選定部1901、集計部1902とから構成され、集計部はさらに実際に表示する内容に対応した記憶領域を持つ表保持部1903を備えている。

【0154】

項目値選定部1901は、操作者がクロス表の一つの軸として指定した項目名（軸項目名）に対して、文書記憶部402に記憶された文書データから、項目値を順次読み出し、重複のない項目値の集合を作成する。また、集計部1902は、表保持部1903の項目値に対応する領域に数値を加算することにより文書の

集計をおこなう。

【0155】

つぎに、クロス表の出力手順について説明する。図20は、実施の形態1による文書処理装置のクロス表の出力手順を示すフローチャートである。図20のフローチャートにおいて、まず、集計に先立ち、表保持部1903の内容を初期化する（ステップS2001）。

【0156】

つぎに、項目値設定部1901により作成された項目値を、表の項目値ラベルに相当する部分に当てはめ（ステップS2002）、クラスタ番号を表す文字列を、クラスタ番号に相当する部分に当てはめる（ステップS2003）。

【0157】

つぎに、加工処理結果保持部408に保持された項目値に対応する文書IDについて、文書記憶部402に記憶された対応する文書を参照し、その軸項目名に対応する項目値を決定する（ステップS2004）。その後、表保持部1903の対応する領域の内容に1を加算する（ステップS2005）。

【0158】

すべての項目値について上記処理をおこなったか否かを判断し（ステップS2006）、すべての項目値について上記処理がおこなわれていない場合（ステップS2006否定）は、ステップS2004へ移行し、ステップS2004～S2006の処理を繰り返しあなう。

【0159】

ステップS2006において、すべての項目値について上記処理がおこなわれた場合（ステップS2006肯定）は、最右列に表示するための行の合計を計算し（ステップS2007）、併せて、最下行に表示するための欄の合計を計算する（ステップS2008）。

【0160】

その後、表保持部1903に構成された表を、順次読み出して出力し（ステップS2009）、すべての処理を終了する。

【0161】

なお、加工処理部405から出力されたデータを、文書記憶部402に送り、文書記憶部402に他のデータとともに記憶するように構成してもよい。文書記憶部402に記憶された加工処理部405から出力されたデータは、出力部406によって表の新たな列として表示することができる。また、表の既存の列を消去し、消去した列へ上書きするようにしてもよい。

【0162】

この構成では、処理の結果である加工処理部405から出力されたデータは、文書記憶部402において、今回の加工処理の対象とならなかった他のデータと対等に取り扱うことができ、その後の分析作業等で、もともとの入力データに存在していたか、分析作業の途中で加工処理によって生成されたのかを区別することなく、加工処理の対象として選択することが可能である。

【0163】

したがって、データの性質や、おこないたい情報分析作業の内容に応じて柔軟に加工処理対象と加工処理内容を選択することができるので、多彩で高度な情報分析作業が可能となる。

【0164】

また、加工処理部405への入力データとして、特徴抽出部404から出力されたデータだけではなく、選択部403により選択されたデータも含めることができる。これにより、文字列の特徴抽出を必要としないデータや、加工処理結果の数値に対してもさらなる加工処理を施すことができるので、より多彩で高度な情報分析が可能となる。

【0165】

図21～図24は、実施の形態1による文書処理装置の出力部406による画面表示の別の例を示す説明図である。図21において、「番号」、「受付日」、「営業所」、「車種」、「年式」、「内容」の他に、分類処理により得られた結果である「クラスタ番号」2101が表示されている。

【0166】

さらに、図21においては、選択部403により「クラスタ番号」2101が選択されており、「クラスタ番号」2101に関するデータが反転表示されてい

る。選択された「クラスタ番号」2101をキーとして、加工処理部405により並べ替え処理をおこなうよう指示をする。

【0167】

並べ替え処理の指示により、並べ替え処理がおこなわれた結果を表示しているのが図22である。図22においては、「クラスタ番号」が「1」のものが集まって表示されるように並べ替えられ、それに続き、「クラスタ番号」が「2」のものが集まって表示されるように並び替えられる。

【0168】

具体的には、「クラスタ番号」が「1」である「番号」が「2」、「11」、「15」、「23」、「35」、「54」、「63」、「73」、「82」の順で並べ替えられ、それに続き「クラスタ番号」が「2」である「番号」が「14」、「18」、「22」、「27」、「37」、...が表示されていることがわかる。

【0169】

つぎに、項目「車種」の欄で、「クラスタ番号」が「1」に属するものを選択する。図23においては、項目「車種」の欄で、「クラスタ番号」が「1」に属するものが選択され、その選択領域2301が反転表示されていることを示している。このように、すでに「クラスタ番号」により並べ替えがおこなわれており、同一クラスタに属するものが集まって表示されているので、画面上の連続した領域として容易に選択することができる。

【0170】

つぎに、選択領域2301について車種別の発生頻度の棒グラフを表示させたのが、図24である。図24において、棒グラフ表示領域2401には、選択領域2301によって選択された「クラスタ番号」が「1」である9つの文書が選択され、その9つの文書を車種別に棒グラフ化したものが表示される。

【0171】

このように、加工処理の対象を柔軟かつ容易に選択でき、選択された対象について多様な加工処理をおこなうことができ、また、その加工処理結果も次回の加工処理の対象とすることができますので、高度な情報分析作業が可能となる。

【0172】

このように、分類等の文字列の特徴量を抽出して、その特徴量を用いておこなう加工処理を実施した後に多種の加工処理をおこなう例を示したが、事前に多種の処理をおこなうことができるようにしてよい。

【0173】

たとえば、「車種」の項目を選択し、これをキーとして並べ替えをおこなった後、集まつたある車種、たとえば、「ABC1600」に対して分類処理をおこなうこともできる。また、入力部401により入力された文書が誤字等の誤りを含んでいる場合、分類等の文字列の特徴量を抽出して、その特徴量を用いて加工処理をおこなう前に、たとえば、文字列の検索・置換処理をおこなって、誤字を一括して修正し、より好適な結果が得られるようにデータを整えることもできる。

【0174】

図25は、実施の形態1による文書処理装置の文書記憶部402の詳細な構成を示すブロック図である。図25において、文書記憶部402は、設定値記憶部2501および設定値送受信部2502を含んでいる。設定値記憶部2501には、文書を分類する際の分類数等の分類情報記憶部2503をはじめとするさまざまな設定値、すなわち文書処理装置の動作に必要な設定値に関する情報を記憶する記憶部を備えている。これにより設定値に関する情報は、文書情報とともに記憶することができる。

【0175】

また、設定値送受信部2502は、設定値記憶部2501によって記憶された設定値に関する情報を他の情報処理装置へ送信する。また、設定値送受信部2502は、他の情報処理装置からの設定値に関する情報を受信する。設定値送受信部2502により受信された設定値に関する情報は、設定値記憶部2501によって記憶される。

【0176】

記憶された設定値に関する情報は、後に文書を再度読み込んだときに同時に読み込まれ設定値記憶部2501に記憶される。この設定値に関する情報は操作者

が所定の操作をすることにより参照することができたり、以後の処理の際に、再利用することができる。これにより、設定値に関する情報を文書とともに保存・管理する事が可能となるので、設定値に関する情報の紛失を防ぎ、好適な設定値を後に再利用することができる。

【0177】

図26～図28は、実施の形態1による文書処理装置の出力部406による画面表示の別の例を示す説明図である。図26において、まず、操作者が分類をおこなうべき対象である「内容」を表示画面上で選択する。それにより選択領域2601が反転表示される。つぎに、メニュー・バー2603から、分類処理ボタン2603を選択すると、分類処理に必要な分類数、すなわち、対象をいくつに分類するかについての問い合わせ画面2604が表示される。

【0178】

操作者が問い合わせ画面2604において分類数を入力すると、この分類数に関する情報が文書記憶部402に記憶される。図26においては、分類数として「50」が入力されたことを示している。

【0179】

その後、操作者が情報分析作業を完了して、メニュー・バー2603のファイルボタン2605の選択によりポップアップする図示を省略する保存ボタンを押下すると、文書記憶部402により、操作者が指示したファイル名が付与され、文書の情報、分類結果とともに記憶される。

【0180】

図27において、分類結果を表示する欄2701にマウスポインタ2702を移動させ、マウスボタンを押下すると、その分類をおこなうことに用いた分類に関する情報および分類設定値に関する情報を表示する分類情報表示画面2703が表示される。これにより、用いた設定値の関連づけが容易に把握することができる。

【0181】

分類情報表示画面2703には、たとえば、分類に関する情報として分類がおこなわれた日時に関する情報を示す「分類日時」、分類の対象となった文書数に

関する情報を示す「分類対象数」等が表示され、また、分類設定値に関する情報として、いくつに分類したかを示す「分類数」、どの品詞に基づいて分類したかを示す「分類品詞」等が表示される。

【0182】

分類処理を実行するたびに新規な表が作成される。図28は、分類結果1を得た後、再度分類処理がおこなわれ、分類結果2が表示された状態を示している。分類結果1を再度表示させたい場合は、画面左下部のラベル上の選択領域2801へマウスポインタを移動させ、マウスボタンを押下する。これにより、分類結果1が再度表示される。その後、分類結果2を再度表示させる場合も同様の操作によりおこなうことができる。

【0183】

また、図28において、各分類処理の実行に用いた設定値に関する情報が対応する表の所定の表示領域2802に表示される。この表示領域2802は、分類結果の表示を隠さないように表示させることができ、また、その表示位置を移動することもできる。これにより、分類結果と、それに用いた設定値の関連づけが容易に把握できる。

【0184】

つぎに、実施の形態1における文書処理装置の文書処理の一連の手順について説明する。図29は、実施の形態1による文書処理装置の文書処理の一連の手順を示すフローチャートである。

【0185】

図29のフローチャートにおいて、まず、文書データが文書処理装置に入力されたか否かを判断する（ステップS2901）。ここで、文書データが入力されるのを待って、文書データが入力された場合（ステップS2901肯定）は、入力された文書データを記憶する（ステップS2902）。なお、ステップS2901およびS2902の各ステップは、文書の入力があるごとに他のステップとは独自におこなわれるようにしてよい。

【0186】

つぎに、記憶された文書データの全部または一部が選択されたか否かを判断す

る（ステップS2903）。ここで、文書データの全部または一部が選択されるのを待って、選択された場合（ステップS2903肯定）は、選択された文書データの全部または一部の文字列の特徴に関するデータの抽出をおこなう（ステップS2904）。

【0187】

その後、ステップS2904において、抽出された文字列の特徴に関するデータに基づいて、分類処理等、所定の加工処理をおこなう（ステップS2905）。続いて、ステップS2905において加工処理がおこなわれたデータを、表形式に展開する等の出力処理をおこなう（ステップS2906）。

【0188】

さらに、ステップS2905において加工処理されてデータを元の文書データに関連づけして記憶する（ステップS2907）。また、加工処理の設定値等の加工処理の内容に関するデータも併せて記憶する（ステップS2908）。

【0189】

その後、ステップS2905において加工処理されたデータの全部または一部が選択されたか否かを判断し（ステップS2908）、選択されなかった場合（ステップS2909否定）は、ステップS2904へ移行し、以後、ステップS2904～S2909の処理を繰り返しおこなう。一方、ステップS2909において、加工処理されたデータの全部または一部が選択された場合（ステップS2909肯定）は、すべての処理を終了する。

【0190】

なお、実施の形態1で説明した文書処理方法は、あらかじめ用意されたプログラムをパーソナルコンピュータやワークステーション等のコンピュータで実行することにより実現される。このプログラムは、ハードディスク、フロッピーディスク、CD-ROM、MO、DVD等のコンピュータで読み取り可能な記録媒体に記録され、コンピュータによって記録媒体から読み出されることによって実行される。またこのプログラムは、上記記録媒体を介して、または伝送媒体として、インターネット等のネットワークを介して配布することができる。

【0191】

つぎに、実施の形態2～6に係る情報分類装置について説明する。なお、以下説明する実施の形態2～6においては、上記のように多くのノイズを含んだものであるとの解釈に基づいて、一回の文書集合からの話題（内容）抽出と位置づけ、文書分類のためのパラメータ（対象文書集合やクラスタ数、類似度測度、ストップワード等）を変化させながら複数化の分類を実行させ、その結果を保持・統合する手段を設けることで、任意の文書集合にどのような内容が含まれるかを漸次的に収集するものである。

【0192】

【実施の形態2】

この発明の実施の形態2に係る文書分類装置を構成する情報処理システムは、図1に示したように実施の形態1の情報処理システムと同様であるので、その説明は省略する。また、サーバー101およびクライアント102のハードウェア構成についても、図2・図3に示したように実施の形態1と同様であるので、その説明は省略する。

【0193】

つぎに、実施の形態2による文書分類装置の機能的構成について説明する。図30は、実施の形態2による文書分類装置の構成を機能的に示すブロック図である。

【0194】

図30のブロック図において、文書分類装置は、入力部3001と、言語解析部3002と、ベクトル生成部3003と、分類部3004と、分類パラメータ指示部3005と、分類結果記憶部3006と、クラスタ特徴表示部3007と、クラスタ特徴算出部3008と、分類体系記憶部3009と、クラスタ選択指示部3010と、分類体系閲覧操作部3011と、を含む構成である。

【0195】

入力部3001、言語解析部3002、ベクトル生成部3003、分類部3004、分類パラメータ指示部3005、分類結果記憶部3006、クラスタ特徴表示部3007、クラスタ特徴算出部3008、分類体系記憶部3009、クラスタ選択指示部3010、分類体系閲覧操作部3011は、ROM202または

302、RAM203または303、あるいはディスク装置306またはハードディスク316等の記録媒体に記録されたプログラムに記載された命令にしたがってCPU201または301等が命令処理を実行することにより、各部の機能を実現する。

【0196】

ここで、入力部3001は、文書データを入力するものであり、たとえば、キーボード209または311、スキャナ313、OCR機能を備えたスキャナ313、またはネットワーク103を経由して文書や文書群を得ることができるI/F204または309等である。

【0197】

また、入力部3001は、上記以外に、文書データを取得することができるものであれば、それらのすべてを含む。たとえば、文書データがデータベース化されている場合に、そのデータベースが記録された媒体を本実施の形態の文書分類装置に組み入れた場合も文書データの入力とする。

【0198】

また、言語解析部3002は、入力部3001により入力された文書データを解析して言語解析情報を得るものであり、ベクトル生成部3003は、言語解析部3002により得られた言語解析情報に基づいて前記文書データに対する文書特徴ベクトルを生成するものである。

【0199】

また、分類部3004は、ベクトル生成部3003により生成された文書特徴ベクトル間の類似度に基づいて文書を分類し、文書の部分集合を生成するものであり、分類パラメータ指示部3005は、分類パラメータを指示するものであり、たとえば、キーボード209または311、マウス210または312、またはネットワーク103を経由して指示情報を得ることができるI/F204または309等である。

【0200】

また、分類結果記憶部3006は、分類部3004により分類された結果、すなわち、分類された文書の部分集合に関する情報を記憶するものである。また、

クラスタ特徴表示部3007は、クラスタ特徴算出部3008により算出されたクラスタ特徴を表示する。

【0201】

クラスタ特徴算出部3008は、分類部3004により生成された文書の部分集合の特徴であるクラスタ特徴を算出するものである。また、分類体系記憶部3009は、クラスタ特徴算出部3008により算出されたクラスタ特徴を分類体系の構成要素として記憶するものである。また、分類体系記憶部3009は、クラスタ選択指示部3010により選択された文書の部分集合を分類体系の構成要素として記憶するものである。すなわち、クラスタ選択指示部3010により選択されたクラスタに所属する全ての文書もしくは所属する文書の一部を分類体系の構成要素として記憶するものである。

【0202】

クラスタ選択指示部3010は、クラスタ表示部3007により表示された複数のクラスタ特徴の中から所望のクラスタを選択するものである。また、クラスタ選択指示部3010は、前記分類部3004により生成された文書の部分集合の中から所望の部分集合を選択するものである。また、分類体系閲覧操作部3011は、分類体系記憶部3009に記憶されたデータを閲覧したい場合に、その閲覧の操作をおこなうものである。

【0203】

つぎに、文書集合に含まれる話題（内容）を抽出することが重要となる好適な例を、アンケート調査等により得られた自由記述回答の分析場面を想定し、その具体例を用いて説明する。

【0204】

近年、たとえば、インターネット等を介して短期間に数千～数万件の自由記述回答を回収することが可能であり、このような機能を用いて大量のテキスト情報の収集をおこなうことができる。

【0205】

アンケート調査により得られた大量のテキスト情報の収集の例として、「オフィスのネットワーク化による無駄を挙げてください」という質問に対して文書で

答えた一つの回答記述を文書とすると、文書集合（クラスタ）は1件ごとの回答の集合ということになる。

【0206】

ここで、操作者（アンケートの分析者）は、そのニーズの一つとして、意見集合（文書集合）にどのような種類の意見（話題）が含まれており、意見の概略を把握したい場合がある。このようなニーズを満たすべく、話題の抽出を類似する意見のまとまり（分類）により実現し、アンケート結果にどのような種類の意見が含まれているかを抽出する。

【0207】

文書分類は、典型的には大きく分けてつぎの3段階のステップから構成される。第1ステップでは、入力部3001により入力された各文書（意見）について、言語解析部3002が、各文書に含まれる単語（あるいは、特定の連続する文字列）を抽出する。この際、たとえば、形態素形跡等の言語解析アルゴリズムが用いられる。

【0208】

第2ステップでは、抽出された単語を列とし、各文書を行とし、要素を単語の出現頻度とした「単語」×「文書」の行列が生成される。なお、一般的な形態素解析機能と構文解析機能を有する言語解析ツールを用いると単語抽出のほかに、単語の品詞情報、複合語（フレーズ）、構文情報等の同時に取得することができ、こうした情報を上記単語×文書の行列を生成する際、考慮することができる。

【0209】

ベクトル生成部3003は、この「単語」×「文書」の行列に基づいて単語で構成される多次元空間内に各文書をベクトル表現する。これには、以下の方法があり、本実施の形態においては、すべての方法を実装している。

【0210】

- (1) 行列の列成分をそのまま利用する方法、
- (2) 各文書の長さ（文字の数やページ数等）や分類対象全体の文書集合内の各単語の出現頻度を考慮して値の重み付けをする方法、
- (3) 上記行列から文書間の内積行列を算出し、これに特異値分解（たとえば、

因子分析や主成分分析、数量化理論第3類等を利用しておこなわれる)を適用して潜在的意味空間を構成する方法、
等である。

【0211】

また、「Representating Documents Using an Explicit Model of Their Similarities (著者名: Brian T. Bartell, Garrison W. Cottrell, and Richard K. Belew, 論文名: Journal of the American Society for Information Science, 学会名: the American Society for Information Science, ページ: 254-271, Vol. 46 No. 4, 発行年: 1995)」においては、上記潜在的意味空間への変換手法を一般化し、文書間の内積行列に、文書が有するほかの文書への参照情報から生成される共参照情報などを付加した行列を用いて、これらの類似性を反映する空間へ文書や単語を射影するための表現空間変換関数を導出しているものもあり、この方法も利用することができる。

【0212】

第3ステップでは、分類部3004が、文書特徴ベクトルの類似度を用いて文書を分類する。具体的には分類対象データに対してカイ自乗法の手法、判別分析の方法、クラスタリングの方法等を適用することにより分類が実行される。

【0213】

また、類似度としては、内積や余弦、ユークリッド距離、マハラノビスの距離等が考えられ、本実施の形態においては、いずれの方法を用いてもよい。

【0214】

また、クラスタリングのアルゴリズムに関するもさまざまなものが公知になっている。クラスタリングは、大別して階層型クラスタリングと非階層型クラスタリングが考えられるが、本実施の形態においては、いずれの方法を用いてもよい。

【0215】

また、分類パラメータ指示部3005は、分類部3004が文書特徴ベクトルを分類するための分類パラメータを指示する。分類部3004は、分類パラメータ指示部3005により指示された分類パラメータにしたがって内部に保持される文書特徴ベクトルを分類する。

【0216】

このようにして、第1ステップ～第3ステップの各処理を実行することにより第1回目の文書分類が終了すると、分類結果は分類結果記憶部3006により保持される。

【0217】

引き続き、クラスタ特徴算出部3008が、分類結果がどのようなクラスタを得ることができたのかを示す特徴、すなわちクラス特徴を算出する。典型的には各クラスタに所属する文書、あるいはその文書の一部を算出するが、その際、クラスタの重心との類似度に基づいて文書をソーティングして出力する。

【0218】

そのほか、クラスタ内で最頻の単語、クラスタに所属する文書数、クラスタ内での文書のばらつきの程度を表すクラスタ内の標準偏差のような数値をクラスタの特徴を表現するものとして算出する。

【0219】

これらのクラスタの特徴情報は、操作者に対して出力（表示）されたクラスタがどのようなもの（どのような特徴を有するもの）かを把握するために算出されるものであり、操作者に対してクラスタの特徴を示すものであれば、上記の内容（特徴）以外のものであってもよい。

【0220】

また、クラスタ特徴算出部3008は、上記のようにクラスタの特徴を示すもの以外に、クラスタ間の関係を示す情報も算出する。階層型クラスタリングの場合は、その上位あるいは下位のクラスタを、非階層型クラスタリングの場合は、クラスタ重心間の類似度に基づく近接のクラスタを算出する。

【0221】

つぎに、クラスタ特徴表示部3007によるクラスタ特徴の表示およびクラスタ選択の内容について説明する。図31は、実施の形態2による文書分類装置のクラスタ特徴表示部3007の表示の一例を示す説明図である。

【0222】

図31において、クラスタ単位で操作者ができるようになっており、各クラスタは「クラスタID」欄3101、「メンバー数」欄3102、「頻度の高い単語」欄3103、「文書内容」欄3104、「重心との類似度」欄3105等の項目から構成される。

【0223】

「クラスタID」欄3101には、クラスタのIDを示す番号が通し番号で付与され、表示される。「メンバー数」欄3102はクラスタに所属する文書あるいは文書の一部の数が算出され、表示される。その中で頻度の高い単語が抽出され「頻度の高い単語」欄3103に表示される。「文書内容」欄3104には文書の内容が表示され、「重心との類似度」欄3105には、数値化された重心との類似度が表示される。これにより、操作者の理解容易性が向上する。

【0224】

操作者は、表示された情報（特徴量）に基づいてクラスタについてその特徴を把握することができる。ここで、内容（特徴）が理解可能なクラスタが一つでもあれば、操作者はクラスタ選択指示部3010によりクラスタを選択することができる。

【0225】

より具体的には、マウス210または312等によって、表示されているクラスタの所定の位置、たとえば、「クラスタID」欄3101へカーソル3110を移動させ、その位置でクリックすることにより、当該クラスタIDのクラスタ全体を選択することができる。なお、選択したクラスタに所属する文書は必ずすべてが選択されるわけではなく、その一部の文書が選択されるようにしてもよい。

【0226】

図31においては、「クラスタID」欄3101がクリックされ、これにより

、クラスタ全体が反転表示しており、当該クラスタ（クラスタID「1」）が選択されたことを示している。

【0227】

また、操作者は、内容が理解可能であるクラスタが存在しない場合は、分類パラメータ指示部3005により分類パラメータの再設定をおこない、再度分類実行をおこなうことができる。

【0228】

クラスタ選択指示部3010により選択されたクラスタIDに関するデータは分類体系記憶部3009へ送信される。分類体系記憶部3009は、このクラスタIDに関するデータに基づいてクラスタ特徴算出部3008からクラスタに関する上記特徴量を検索し記憶する。

【0229】

また、分類体系記憶部3009は、同様に、分類結果記憶部3006から分類結果を検索し記憶する。さらに、分類体系記憶部3009は、操作者により入力されたクラスタに関するコメント（たとえば、「ネットワークの維持費が高い」等）の情報を併せて記憶することもできる。このように、操作者が作成した情報を分類体系の構成要素として記憶することにより、分類体系の利用価値がより向上する。

【0230】

なお、分類体系記憶部3009により記憶されたデータは、別途閲覧操作用のインターフェイスを設けることにより、選択・保持するクラスタの内容の閲覧や、クラスタ間の意味的な関連を手動あるいは、保持されているクラスタ重心間の類似度等を用いて自動で、構造化・体系化することができる。

【0231】

つぎに、実施の形態2の文書分類装置の一連の処理の手順について説明する。図32は、実施の形態2による文書分類装置の一連の処理の手順を示すフローチャートである。図32のフローチャートにおいて、まず、分類の対象となる文書が入力される（ステップS3201）。

【0232】

つぎに、入力された文書の言語が解析され（ステップS3202）、解析された結果、すなわち、抽出された単語に基づいて、文書特徴ベクトルが生成される（ステップS3203）。

【0233】

その後、分類パラメータの指示があるのを待って、分類パラメータの指示があった場合（ステップS3204肯定）は、指示があった分類パラメータにしたがって文書を分類し（ステップS3205）、その結果、すなわち、クラスタに関する情報を記憶する（ステップS3206）。

【0234】

つぎに、分類されたクラスタの特徴を算出し（ステップS3207）、算出された結果を表示する（ステップS3208）。表示されたクラスタの中から、クラスタが選択されたか否かを判断し（ステップS3209）、選択されなかった場合（ステップS3209否定）は、ステップS3204へ移行し、再度分類パラメータの指示があるのを待つ（ステップS3204）。

【0235】

一方、ステップS3209において、クラスタが選択された場合（ステップS3209肯定）は、選択されたクラスタに関して分類体系を生成し、記憶する（ステップS3210）。この際、操作者により入力されたクラスタに関する情報を併せて記憶することもできる。これにより、一連の処理を終了する。

【0236】

以上説明したように、実施の形態2による文書分類装置によれば、分類対象である文書群での文書間の類似性に基づいて、各文書をそれら文書間の意味的な関連性を反映しうる表現空間へ変換するための表現空間変換関数を算出し、その表現空間で文書分類をおこなうことにより、操作者の意図を反映しうる文書分類を実現することができる。

【0237】

したがって、分類部3004によりクラスタを得ることができるとともに、クラスタ特徴算出部3008・分類体系記憶部3009により、クラスタ重心間の類似度等を用いて、クラスタの内容に基づくクラスタの構造化・体系化をおこな

うことができる。

【0238】

また、クラスタ選択指示部3010により選択されたクラスタのみを用いて、より操作者の意図したものに近いクラスタの構造化・体系化をおこなうことができる。

【0239】

〔実施の形態3〕

さて、上述した実施の形態2に加えて、以下に説明する実施の形態3のように、さらにベクトル記憶部と、ベクトル修正部とを含む構成とするようにしてもよい。

【0240】

実施の形態3による文書分類装置を構成する情報処理システムは、図1に示したように実施の形態1と同様であるので、その説明は省略する。また、サーバー101およびクライアント102のハードウェア構成についても、図2・図3に示したように実施の形態1と同様であるので、その説明は省略する。

【0241】

つぎに、実施の形態3による文書分類装置の機能的構成について説明する。図33は、この発明の実施の形態3による文書分類装置の構成を機能的に示すブロック図である。図33において、実施の形態2の図30と同一のものに関しては同じ符号を付して、その説明を省略する。

【0242】

図33のブロック図において、文書分類装置は、入力部3001、言語解析部3002、ベクトル生成部3003、分類部3004、分類パラメータ指示部3005、分類結果記憶部3006、クラスタ特徴表示部3007、クラスタ特徴算出部3008、分類体系記憶部3009、クラスタ選択指示部3010、分類体系閲覧操作部3011のほか、ベクトル記憶部3301と、ベクトル修正部302とを含む構成である。

【0243】

ベクトル記憶部3301は、ベクトル生成部3003により生成された文書特

文書特徴ベクトルを記憶するものである。また、ベクトル修正部3302は、文書特徴ベクトル記憶部3301により記憶された文書特徴ベクトルを、クラスタ選択指示部3010により選択された部分集合に属する文書の文書特徴ベクトルを除去したのこりとなるように修正するものである。

【0244】

また、分類部3004は、ベクトル修正部3302により修正された文書特徴ベクトルに基づいて文書を分類する。

【0245】

なお、ベクトル記憶部3301、ベクトル修正部3302は、ROM202または302、RAM203または303、あるいはディスク装置306またはハードディスク316等の記録媒体に記録されたプログラムに記載された命令にしたがってCPU201または301等が命令処理を実行することにより、各部の機能を実現する。

【0246】

ベクトル生成部3003において生成された文書特徴ベクトル（列ベクトル）・単語（単語特徴）ベクトル（行ベクトル）はベクトル記憶部3301によって記憶される。これは、次回以降の分類実行の際に利用する文書特徴ベクトルを確保するためである。

【0247】

ベクトル修正部3302は、クラスタ選択指示部3010により選択されたクラスタに所属する文書のすべてあるいはその一部の文書を除き、次回以降もこれらの文書が除かれるよう削除する。削除された文書特徴ベクトルはベクトル記憶部3301により記憶される。

【0248】

この結果、ベクトル記憶部3301に記憶されているベクトルデータのうち、選択されたクラスタに所属する文書（もしくは操作者に指定されたその一部）列ベクトルを除いたものが、次回以降の分類が実行される際に利用されるデータとなる。

【0249】

つぎに、実施の形態3の文書分類装置の一連の処理の手順について説明する。

図34は、実施の形態3による文書分類装置の一連の処理の手順を示すフローチャートである。図2のフローチャートにおいて、まず、分類の対象となる文書が入力される（ステップS3401）。

【0250】

つぎに、入力された文書の言語が解析され（ステップS3402）、解析された結果、すなわち、抽出された単語に基づいて、文書特徴ベクトルが生成され（ステップS3403）、生成された文書特徴ベクトルが記憶される（ステップS3404）。

【0251】

その後、分類パラメータの指示があるのを待って、分類パラメータの指示があった場合（ステップS3405肯定）は、指示があった分類パラメータにしたがって文書を分類し（ステップS3406）、その結果、すなわち、クラスタに関する情報を記憶する（ステップS3407）。

【0252】

つぎに、分類されたクラスタの特徴を算出し（ステップS3408）、算出された結果を表示する（ステップS3409）。表示されたクラスタの中から、クラスタが選択されたか否かを判断し（ステップS3410）、選択されなかった場合（ステップS3410否定）は、ステップS3405へ移行し、再度分類パラメータの指示があるのを待つ（ステップS3405）。

【0253】

一方、ステップS3410において、クラスタが選択された場合（ステップS3410肯定）は、選択されたクラスタに関して分類体系を生成し、記憶する（ステップS3411）。この際、操作者により入力されたクラスタに関する情報を併せて記憶することもできる。その後、繰り返し処理をおこなう旨の指示があったか否かを判断する（ステップS3412）。

【0254】

ステップS3412において、繰り返して処理をおこなう旨の指示があった場合（ステップS3412肯定）は、選択されたクラスタに所属する文書のすべて

あるいはその一部の文書を除くように文書特徴ベクトルを修正する（ステップS3413）。その後、ステップS3405へ移行し、以後、ステップS3405～S3413の各処理を繰り返しおこなう。

【0255】

一方、ステップS3412において、繰り返して処理をおこなう旨の指示がない場合（ステップS3412否定）は、これにより、一連の処理をすべて終了する。

【0256】

以上説明したように、実施の形態3による文書分類装置によれば、ベクトル修正部3301により、既知になったクラスタの影響を排除した新たなクラスタを生成することができる。

【0257】

〔実施の形態4〕

さて、上述した実施の形態3においては、ベクトル記憶部およびベクトル修正部とを含む構成であったが、以下に説明する実施の形態4のように、ベクトル修正部に代わりに、文書表現空間修正部を含む構成とするようにしてもよい。

【0258】

実施の形態4による文書分類装置を構成する情報処理システムは、図1に示したように実施の形態1と同様であるので、その説明は省略する。また、サーバー101およびクライアント102のハードウェア構成についても、図2・図3に示したように実施の形態1と同様であるので、その説明は省略する。

【0259】

つぎに、実施の形態4による文書分類装置の機能的構成について説明する。図35は、この発明の実施の形態4による文書分類装置の構成を機能的に示すブロック図である。図35において、実施の形態2の図30と同一のものに関しては同じ符号を付して、その説明を省略する。

【0260】

図35のブロック図において、文書分類装置は、入力部3001、言語解析部3002、ベクトル生成部3003、分類部3004、分類パラメータ指示部3

005、分類結果記憶部3006、クラスタ特徴表示部3007、クラスタ特徴算出部3008、分類体系記憶部3009、クラスタ選択指示部3010、分類体系閲覧操作部3011のほか、ベクトル記憶部3501と、文書表現空間修正部3502とを含む構成である。

【0261】

ベクトル記憶部3501は、ベクトル生成部3003により生成された文書特徴ベクトルを記憶するものである。また、文書表現空間修正部3502は、文書特徴ベクトル記憶部3501により記憶された文書特徴ベクトル間の類似度を判断する際の文書表現空間を前記クラスタ選択指示部3010により選択された部分集合から算出する特微量に基づいて修正するものである。

【0262】

また、分類部3004は、文書表現空間修正部3502により修正された文書表現空間を用いて、ベクトル生成部3003により生成された文書特徴ベクトル間の類似度に基づいて文書を分類する。

【0263】

なお、ベクトル記憶部3501、文書表現空間修正部3502は、ROM202または302、RAM203または303、あるいはディスク装置306またはハードディスク316等の記録媒体に記録されたプログラムに記載された命令にしたがってCPU201または301等が命令処理を実行することにより、各部の機能を実現する。

【0264】

つぎに、文書表現空間修正部3502の内容について説明する。実施の形態3におけるベクトル修正部3302にあっては、既知になったクラスタの影響を排除するために文書特徴ベクトルを除去するが、文書特徴ベクトルを表現する多次元空間自体の変更はおこなわれない。

【0265】

したがって、前回の分類実行の結果、操作者により選択されたクラスタの形成特徴を次回の分類実行の際に排除したい場合は、文書ベクトルを表現する空間自体の変更が必要となる。

【0266】

そこで、文書表現空間修正部3502を備え、文書表現空間の修正をおこなうものである。ここで、文書表現空間の特徴次元を変更する例として、操作者により選択されたクラスタの重心と類似度の高い特徴次元の削除をおこなうことについて説明する。

【0267】

操作者により選択されたクラスタの重心はベクトルとして表現することができるので、このクラスタ重心ベクトルとベクトル記憶部3501に記憶されている文書表現空間の各特徴次元との類似度を算出することにより、類似度の高い特徴次元を判別する。

【0268】

なお、類似の測度としては、余弦、内積、ユークリッド距離、マハラノビス距離等を用いる。また、判別に関してはある類似度以上を削除対象として採用するようなしきい値処理による判別や、類似度の高い順にある一定数を削除対象として採用する定数処理による判別を用いる。また、判別分析等も用いることができる。

【0269】

文書表現空間修正部3502は、上述のような削除対象の特徴次元を算出して、特徴次元の削除をおこなう。特徴次元の削除は、ベクトル記憶部3501に記憶されている「特徴次元（単語）」×「文書」の行列から判別された特徴次元について行ベクトルを削除することによりおこなう。文書表現空間修正部3502により修正された文書ベクトルは、次回以降の分類のために、ベクトル記憶部3501に記憶される。

【0270】

つぎに、実施の形態4の文書分類装置の一連の処理の手順について説明する。図36は、実施の形態4による文書分類装置の一連の処理の手順を示すフローチャートである。図36のフローチャートにおいて、まず、分類の対象となる文書が入力される（ステップS3601）。

【0271】

つぎに、入力された文書の言語が解析され（ステップS3602）、解析された結果、すなわち、抽出された単語に基づいて、文書特徴ベクトルが生成され（ステップS3603）、生成された文書特徴ベクトルが記憶される（ステップS3604）。

【0272】

その後、分類パラメータの指示があるのを待って、分類パラメータの指示があった場合（ステップS3605肯定）は、指示があった分類パラメータにしたがって文書を分類し（ステップS3606）、その結果、すなわち、クラスタに関する情報を記憶する（ステップS3607）。

【0273】

つぎに、分類されたクラスタの特徴を算出し（ステップS3608）、算出された結果を表示する（ステップS3609）。表示されたクラスタの中から、クラスタが選択されたか否かを判断し（ステップS3610）、選択されなかった場合（ステップS3610否定）は、ステップS3605へ移行し、再度分類パラメータの指示があるのを待つ（ステップS3605）。

【0274】

一方、ステップS3610において、クラスタが選択された場合（ステップS3610肯定）は、選択されたクラスタに関して分類体系を生成し、記憶する（ステップ3611）。この際、操作者により入力されたクラスタに関する情報を併せて記憶することもできる。その後、繰り返し処理をおこなう旨の指示があったか否かを判断する（ステップS3612）。

【0275】

ステップS3612において、繰り返して処理をおこなう旨の指示があった場合（ステップS3612肯定）は、「特徴次元（単語）」×「文書」の行列から判別された特徴次元について行ベクトルを削除することにより文書表現空間を修正する（ステップS3613）。その後、ステップS3605へ移行し、以後、ステップS3605～S3613の各処理を繰り返しおこなう。

【0276】

一方、ステップS3612において、繰り返して処理をおこなう旨の指示がな

かった場合（ステップS3612否定）は、これにより、一連の処理を終了する。

【0277】

以上説明したように、実施の形態4による文書分類装置によれば、前回の分類実行の結果、文書表現空間修正部3502により操作者に選択されたクラスタの形成特徴を次回の分類実行時に排除することができ、排除した状態で新たなクラスタを生成することができる。

【0278】

【実施の形態5】

さて、上述した実施の形態3または実施の形態4においては、ベクトル修正部または文書表現空間修正部のいずれか一方のみを含む構成であったが、以下に説明する実施の形態5のように、ベクトル修正部および文書表現空間修正部の両方を含む構成とするようにしてもよい。

【0279】

実施の形態5による文書分類装置を構成する情報処理システムは、図1に示したように実施の形態1と同様であるので、その説明は省略する。また、サーバー101およびクライアント102のハードウェア構成についても、図2・図3に示したように実施の形態1と同様であるので、その説明は省略する。

【0280】

つぎに、実施の形態5による文書分類装置の機能的構成について説明する。図37は、この発明の実施の形態5による文書分類装置の構成を機能的に示すブロック図である。図37において、実施の形態2の図30と同一のものに関しては同じ符号を付して、その説明を省略する。

【0281】

図37のブロック図において、文書分類装置は、入力部3001、言語解析部3002、ベクトル生成部3003、分類部3004、分類パラメータ指示部3005、分類結果記憶部3006、クラスタ特徴表示部3007、クラスタ特徴算出部3008、分類体系記憶部3009、クラスタ選択指示部3010、分類体系閲覧操作部3011のほか、ベクトル記憶部3701と、ベクトル修正部3

702と、文書表現空間修正部3703とを含む構成である。

【0282】

ベクトル記憶部3701は、ベクトル生成部3003により生成された文書特徴ベクトルを記憶するものである。また、ベクトル修正部3702は、文書特徴ベクトル記憶部3701により記憶された文書特徴ベクトルを分類部3004により生成された文書の部分集合の文書特徴ベクトルを除去したのこりの文書特徴ベクトルとなるように修正するものである。

【0283】

また、文書表現空間修正部3703は、ベクトル記憶部3701により記憶された文書特徴ベクトル間の類似度を判断する際の文書表現空間を前記クラスタ選択指示部3010により選択されたクラスタ特徴に基づいて修正するものである。

【0284】

また、分類部3004は、文書表現空間修正部3703により修正された文書表現空間を用いて、ベクトル修正部3702により修正された文書特徴ベクトル間の類似度に基づいて文書を分類する。

【0285】

なお、ベクトル記憶部3701、ベクトル修正部3702、文書表現空間修正部3703は、ROM202または302、RAM203または303、あるいはディスク装置306またはハードディスク316等の記録媒体に記録されたプログラムに記載された命令にしたがってCPU201または301等が命令処理を実行することにより、各部の機能を実現する。

【0286】

つぎに、ベクトル修正部3702および文書表現空間修正部3703の内容について説明する。実施の形態4においては、選択されたクラスタに所属する文書は次回以降の分類実行の際にも使用される。

【0287】

実施の形態5では、ベクトル修正部3702および文書表現空間修正部3703の両方を具備することにより、選択されたクラスタに所属する文書を次回の分

類実行の際に除去し、次回の分類実行の際には分類対象文書としないようにする。

【0288】

実施の形態4においては、話題抽出の側面を強調し、ある文書が複数の話題として分類される可能性を前提としており、たとえば、ネットワーク化に関する調査における「エンドユーザーがソフトウェアのインストール方法について聞いてくるのでシステム管理者としての仕事ができない」という回答について言えば、この意見は「ソフトウェアの操作方法理解に関する困難性」という話題として分類され得るし、「システム管理者の仕事の多忙さ」という話題で分類される可能性もある。

【0289】

実施の形態4においては、いずれにしても、「ソフトウェアの操作方法理解に関する困難性」というクラスタと「システム管理者の仕事の多忙さ」というクラスタの両方とも抽出したいというニーズに応えている。

【0290】

これとは反対に、操作者は、一度抽出した話題は既知であるので、次回の分類の際はなるべく異なる分類結果が欲しいとするケースも考えられる。実施の形態5では、このような要求に応えるため、ベクトル修正部3702により、n回目の分類で選択されたクラスタに所属する文書のすべてまたはその一部を次回以降の分類を実行する際、分類対象から除去するものである。

【0291】

クラスタ選択指示部3010により選択指示を受けたクラスタの所属文書はベクトル記憶部3701において列ベクトルの形式で記憶されているため、ベクトル修正部3702では劣ベクトルを除去することで、次回以降の分類実行用の分類対象文書集合を生成する。

【0292】

さらに、実施の形態4と同様に、選択されたクラスタにより文書表現空間修正部3703は、ベクトル記憶部3701に記憶されている行列から特徴次元を削除する。

【0293】

つぎに、実施の形態5の文書分類装置の一連の処理の手順について説明する。図38は、実施の形態5による文書分類装置の一連の処理の手順を示すフローチャートである。図38のフローチャートにおいて、まず、分類の対象となる文書が入力される（ステップS3801）。

【0294】

つぎに、入力された文書の言語が解析され（ステップS3802）、解析された結果、すなわち、抽出された単語に基づいて、文書特徴ベクトルが生成され（ステップS3803）、生成された文書特徴ベクトルが記憶される（ステップS3804）。

【0295】

その後、分類パラメータの指示があるのを待って、分類パラメータの指示があった場合（ステップS3805肯定）は、指示があった分類パラメータにしたがって文書を分類し（ステップS3806）、その結果、すなわち、クラスタに関する情報を記憶する（ステップS3807）。

【0296】

つぎに、分類されたクラスタの特徴を算出し（ステップS3808）、算出された結果を表示する（ステップS3809）。表示されたクラスタの中から、クラスタが選択されたか否かを判断し（ステップS3810）、選択されなかった場合（ステップS3810否定）は、ステップS3805へ移行し、再度分類パラメータの指示があるのを待つ（ステップS3805）。

【0297】

一方、ステップS3810において、クラスタが選択された場合（ステップS3810肯定）は、選択されたクラスタに関して分類体系を生成し、記憶する（ステップS3811）。この際、操作者により入力されたクラスタに関する情報を併せて記憶することもできる。その後、繰り返し処理をおこなう旨の指示があったか否かを判断する（ステップS3812）。

【0298】

ステップS3812において、繰り返して処理をおこなう旨の指示があった場

合（ステップS3812肯定）は、選択されたクラスタに所属する文書のすべてあるいはその一部の文書を除くように文書特徴ベクトルを修正する（ステップS3813）。

【0299】

ステップS3813に引き続き、「特徴次元（単語）」×「文書」の行列から判別された特徴次元について行ベクトルを削除することにより文書表現空間を修正する（ステップS3814）。その後、ステップS3805へ移行し、以後、ステップS3805～S3814を繰り返しおこなう。

【0300】

一方、ステップS3812において、繰り返して処理をおこなう旨に指示がない場合（ステップS3812否定）は、これにより、一連の処理をすべて終了する。

【0301】

以上説明したように、実施の形態5による文書分類装置によれば、ベクトル修正部3702が、既知になったクラスタの影響を排除し、かつ、文書表現空間修正部3703が、前回の分類実行の結果、操作者に選択されたクラスタの形成特徴を次回の分類実行時に排除することができ、排除した状態で新たなクラスタを生成することができる。

【0302】

〔実施の形態6〕

さて、上述した実施の形態2または実施の形態4においては、繰り返し分類処理をおこなった場合に、ある文書が何度選択されたかその情報については考慮していなかったが以下に説明する実施の形態6のように、選択情報付与部を含む構成とし、選択情報をクラスタ特徴とともに表示するようにしてもよい。

【0303】

実施の形態6による文書分類装置を構成する情報処理システムは、図1に示したように実施の形態1と同様であるので、その説明は省略する。また、サーバー101およびクライアント102のハードウェア構成についても、図2・図3に示したように実施の形態1と同様であるので、その説明は省略する。

【0304】

つぎに、実施の形態6による文書分類装置の機能的構成について説明する。図39は、この発明の実施の形態6による文書分類装置の構成を機能的に示すブロック図である。図39において、実施の形態4の図35と同一のものに関しては同じ符号を付して、その説明を省略する。

【0305】

図39のブロック図において、文書分類装置は、入力部3001、言語解析部3002、ベクトル生成部3003、分類部3004、分類パラメータ指示部3005、分類結果記憶部3006、クラスタ特徴表示部3007、クラスタ特徴算出部3008、分類体系記憶部3009、クラスタ選択指示部3010、分類体系閲覧操作部3011、ベクトル記憶部3501、文書表現空間修正部3502のほか、選択情報付与部3901を含む構成である。

【0306】

選択情報付与部3901は、分類部3004により生成された文書の部分集合に所属する文書のすべてあるいは一部が選択された場合に選択されたことを示す選択情報を付与する。また、クラスタ特徴表示部3007は、クラスタ特徴を表示するとともに、選択情報付与部3901により付与された選択情報を表示する。

【0307】

なお、選択情報付与部3901は、ROM202または302、RAM203または303、あるいはディスク装置306またはハードディスク316等の記録媒体に記録されたプログラムに記載された命令にしたがってCPU201または301等が命令処理を実行することにより、機能を実現する。

【0308】

つぎに、選択情報付与部3901の詳細な内容について説明する。アンケートの調査の例において、独自性の高いユニークな意見は貴重であることが経験的に知られている。これは、調査を企画する担当者が予想できなかった意見である場合が多いからである。

【0309】

そこで、操作者に選択されたクラスタに所属する文書を、次回以降の分類実行の際に使用する場合において、クラスタ特徴表示部3007で個々の文書を表示する際に、各文書が何回選択されたかを示すことで、多重に利用される文書の識別性を向上させ、かつ一度も選択されない文書の識別性も向上させることができ

る。

【0310】

図40は、実施の形態6による文書分類装置の分類結果記憶部3006において設けられたテーブル4000を示す説明図である。図40において、文書IDごとにテーブル化されており、テーブル4000は、各文書がどのサイクルに分類実行の際に操作者に選択されたかを記録する。すなわち、選択された場合は選択情報として「1」を記録し、選択されなかった場合は選択情報として「0」を記録する。

【0311】

たとえば、4回分類が実行された際、文書IDの「1」、第1回目および第2回目の分類実行時に操作者に選択されたことを示し、第3回目、第4回目の分類実行時には選択されなかったことを示している。一方、文書IDの「2」は、未だ一度も選択されておらず、操作者にとって未知の意見という可能性を示唆している。

【0312】

こうした情報に基づいて、クラスタ特徴表示部3007が文書を操作者に表示する際、たとえば、選択された回数に応じて表示を変化させるようにするとよい。変化させる視覚的特性としては、たとえば文字や背景の色の濃度や彩度等が考えられる。

【0313】

また、直接的に数字やグラフ等で選択された回数を表現することもできる。いずれにしてもよ選択される文書と一度も選択されていない文書とを視覚的に識別できる表示形式であれば、上記のものに限らない。

【0314】

また、上記選択情報を分類体系閲覧操作部3011の閲覧操作により閲覧でき

るようにしてよい。

【0315】

つぎに、選択情報付与部3901の処理の内容について説明する。図41は、実施の形態6による文書分類装置の選択情報付与部3901の処理の手順を示すフローチャートである。図41のフローチャートにおいて、まず、分類処理がおこなわれ（ステップS4101）、それに引き続き、最初の文書が抽出される（ステップS4102）。

【0316】

抽出された文書が、ステップS4101における分類処理の際に選択されたか否かを判断する（ステップS4103）。ここで、選択された場合（ステップS4103肯定）は、選択情報としてデータ「1」を記録する（ステップS4104）。一方、選択されなかった場合（ステップS4103否定）は、選択情報としてデータ「0」を記録する（ステップS4105）。

【0317】

つぎに、すべての文書について処理が終了したか否かを判断する（ステップS4106）。ここで、すべての文書について処理が終了していない場合（ステップS4106否定）は、つぎに文書を抽出し（ステップS4107）、ステップS4103へ移行し、以後、ステップS4103～S4107を繰り返しおこなう。

【0318】

一方、ステップS4106において、すべての文書について処理が終了した場合（ステップS4106肯定）は、ステップS4101へ移行し、再度分類処理がおこなわれる（ステップS4101）。このようにして、分類処理がおこなわれる回数だけ、ステップS4101～S4107の各処理が繰り返しおこなわれる。

【0319】

以上説明したように、実施の形態6によれば、選択情報付与部3901が選択情報を付与し、その選択情報をクラスタ特徴表示部3007が表示するので、多重に利用される文書の識別性および一度も選択されない文書の識別性を向上させ

ることができる。

【0320】

なお、実施の形態2～5で説明した文書分類方法は、あらかじめ用意されたプログラムをパーソナルコンピュータやワークステーション等のコンピュータで実行することにより実現される。このプログラムは、ハードディスク、フロッピーディスク、CD-ROM、MO、DVD等のコンピュータで読み取り可能な記録媒体に記録され、コンピュータによって記録媒体から読み出されることによって実行される。またこのプログラムは、上記記録媒体を介して、または伝送媒体として、インターネット等のネットワークを介して配布することができる。

【0321】

つぎに、実施の形態7～16に係る情報分類装置について説明する。本発明の実施の形態では、自然言語で記述された一つ以上の文の集まりであり、かつその一つ以上の文の集まりが分類される対象である場合、それを文書と言う。具体的な例をあげれば、IPC分類等により分類される公開特許公報や、政治・経済・文化・科学技術等の特定分野に分類される新聞記事も文書であるし、それらから請求項や特定の一文を取り出したものであっても、請求項という分類に含まれる文であるか、用途等により分類可能な特定の一文であれば文書とみなす。以下、図面によりこの発明の実施の形態7～16を詳細に説明する。

【0322】

〔実施の形態7〕

図42はこの発明の実施の形態7を示す文書分類装置の構成ブロック図である。図42に示したように、実施の形態7の文書分類装置は、文書データ群を入力する文書入力部（文書入力手段）5001、それぞれの文書データを所定の基準に基づいて一つまたは複数の分割文書データに分割する文書分割部（文書分割手段）5002、上記文書データと分割文書データとを対応付けるマップを生成する文書一分割文書対応マップ生成部（文書一分割文書対応マップ生成手段）5003を備えている。

【0323】

また、上記文書分類装置は、分割文書データつまり分割された文書を分類する

分割文書分類部（分割文書分類手段）5004、分割文書分類結果情報を生成する分割文書分類結果生成部（分割文書分類結果生成手段）5005、上記文書一分割文書対応マップと上記分割文書分類結果情報とを用いて上記文書データの分類結果情報を生成する文書分類結果生成部（文書分類結果生成手段）5006などを備えている。

【0324】

なお、上記文書分割部5002、文書一分割文書対応マップ生成部5003、分割文書分類部5004、分割文書分類結果生成部5005、文書分類結果生成部5006は共有または独自のプログラム記憶用メモリおよびプログラムにしたがって動作するC P Uを有している。

【0325】

以下、図4.2などにしたがって、実施の形態7の文書分類装置、文書分類方法を詳細に説明する。まず、文書入力部5001により、文書群が入力される。上記文書入力部5001はキーボード、OCR装置、着脱型記録媒体、またはネットワーク通信手段を備え、それらのいずれか一つを介して文書データ群を入力するのである。

【0326】

そして、文書分割部5002が上記文書データ群を取得し、それぞれの文書データを所定の基準に基づいて分割し、一つの文書データから一つまたは複数の分割文書データを生成する。なお、文書データを分割する方法としては、文書の構造情報や文書を構成する要素情報を用いたり、利用者が指定する方法などを用いるが、ここでは、その方法は問わないこととする。

【0327】

図4.3に、この文書分類装置／文書分類方法でおこなわれる、文書データから複数の分割文書データを生成する一例を示す。この例に示した文書1には複数のニューストピックが記述されており、1日分のトピックが文書単位となっている。図示したように、この文書ではそれぞれのニューストピックが二つの改行コードにより分離されているので、この規則を用いて一つの文書である文書1を分割し、一つが一つのトピックにより形成される分割文書1-1～1-7の7つの分

割文書データを生成する。なお、分割前の文書1も分割文書データとして含めることもできるが、ここでは含めないことにする。

【0328】

文書が分割されると、文書一分割文書対応マップ生成部5003が分割前の文書データとその文書データから生成された分割文書データとを対応付けるマップを生成する。たとえば、個々の文書データを一意に示す識別子と個々の分割文書データを一意に示す識別子とから構成されるマップ、あるいは文書データごとに分割文書データを一意に示す識別子からなるマップを生成するのである。なお、文書データと分割文書データを対応付ける方法についてはここでは問わないこととする。

【0329】

図44に、文書一分割文書対応マップを生成する一例を示す。図44において、文書1～文書3は文書データを示し、分割文書1～分割文書12は分割文書データを示している。図示のように、それぞれの文書データおよび分割文書データにそれを一意に識別することができる識別番号（識別子）を付与し、上記文書データの識別番号と分割文書データの識別番号とを図44の左下に示したテーブル形式で対応づけている。なお、任意の複数の分割文書データが文書分類にて用いられる基準において同一とみなすことができる場合は、それらの識別番号を同一にしてもよい。

【0330】

続いて、分割文書分類部5004が上記分割文書を対象に文書分類をおこなう。個々の分割文書に対して、たとえば、言語処理を施し、文書中に含まれているそれぞれの単語の出現頻度を計数し、それに基づいてそれぞれの文書の特徴を計量的に表す特徴ベクトルを求め、それらの特徴ベクトルに対してカイ自乗法、判別分析手法、またはクラスタ分析手法などを適用することにより文書分類をおこなう。

【0331】

つぎに、図45に示すように、分割文書分類結果生成部5005が上記の分割文書分類の結果に基づいた分割文書分類結果情報を生成する。

【0332】

ここで、分割文書分類結果情報とは、たとえば、各分割文書データの所属カテゴリに関する情報（たとえば、図45に示した「分割文書データを3つのカテゴリに分類した結果」という表中の「分類カテゴリ」および「所属カテゴリの代表値との距離」の項の情報）、生成された所属カテゴリ個々に関する情報（たとえば、図45に示した「分類カテゴリに関する情報」という表中の「代表値」および「所属データ数（分割文書数）」の項の情報）、生成された所属カテゴリ間の情報（たとえば図45に示した「分類カテゴリ間の距離」という表の中の情報）などである。なお、利用者は上記のような種々の情報を分類結果分析の際の基礎データとして利用することができる。

【0333】

図45は、12個の分割文書データをそれらの有する計量的特徴ベクトルを用いて3つのカテゴリに分類した場合の分類結果の生成例である。分割文書データの有する計量的な3次元ベクトル（ベクトルの成分数は分類対象文書群に生起するすべての単語の種類数になるが、ここでは、いくつかの単語が縮退した3次元ベクトルに線形変換している）に対してたとえばクラスタ分析手法の一つであるWard法などを適用することで3つのカテゴリに分類することができる。

【0334】

つまり、各分割文書データは図示したように3つのカテゴリのうちのいずれか一つに属する。なお、所属カテゴリの代表値とは、所属分割文書データの特徴ベクトルの平均値（所属分割文書データの重心）である。

【0335】

また、所属カテゴリの代表値との距離（類似度に対応する）は、たとえば、図45の分割文書3については、分割文書データ特徴ベクトルの項における分割文書3の値と、分割文書3の分類カテゴリであるカテゴリ2の代表値（所属分割文書データの重心）の項の値により、以下の式から求めることができる。

【0336】

$$\sqrt{((3.00 - 2.66)^2 + (2.00 - 2.00)^2 + (4.00 - 3.66)^2)} = 0.48$$

上記の所属カテゴリの代表値との距離が小さいほど、そのカテゴリに属する平均的分割文書との類似度が高いということになる。

【0337】

なお、分割文書分類結果情報としては、図45に示した以外にも、カテゴリ内分散やカテゴリ間分散、各カテゴリにおける類似度のレンジなどさまざまな統計量を生成することができる。

【0338】

続いて、文書分類結果生成部5006が上記文書-分割文書対応マップと上記分割文書分類結果情報とを用いて、たとえば図46に示すような、上記文書データの分類結果情報を生成する。図46の例では、図示したように、各分類カテゴリごとに、所属する分割文書データ、その類似度（所属カテゴリの代表値との距離）、分割文書データの属する分割前文書データ（所属文書）、文書占有率（分割文書データの当該カテゴリに所属する割合）、分割文書データの所属文書における相対位置（順序）、所属カテゴリ内での当該分割文書データの類似度の順位などを生成している。

【0339】

なお、上記において、所属文書は文書-分割文書対応マップから、それ以外の分類結果情報は分割文書分類結果情報から得ている。文書分類結果生成部5006は図46に示した情報以外にも、各カテゴリ内での分散、分割文書データの所属カテゴリ内での偏差値などさまざまな統計量、文書データや分割文書データの内容などを分類結果情報として利用することもできる。

【0340】

また、上記においては、すべての結果を分割文書データを単位とした表形式で表現しているが、分類カテゴリや文書データを単位として表現することもできる。また、分類結果情報をテキスト表現にするだけでなく、グラフィカルな表現にして、利用者が理解しやすいようにすることも可能である。

【0341】

こうして、本実施の形態によれば、一つの文書が分割され、分割文書が分類され、分割前文書と上記分割文書との対応が利用者に示され、上記分割文書の分類

結果が利用者に示されるので、一つの文書の中に複数の話題や意味が含まれている場合に、ある特定の話題や意味に限定されたカテゴリに分類されたり、利用者の意図するカテゴリとは異なるカテゴリに分類されたりすることがなく、したがって、利用者がその分類カテゴリをよく理解できる。また、分割前文書（所属文書）中の分割文書の位置なども示されるので、利用者は文書群中の読みたい部分を効率的に読むことができる。

【0342】

〔実施の形態8〕

図47は本発明の実施の形態8に係る文書分類装置の構成ブロック図である。図示したように、実施の形態8の文書分類装置は、図42に示した実施の形態7の構成に加え、文書データを保存する文書保存部（文書保存手段）5007、分割文書データを保存する分割文書保存部（分割文書保存手段）5008、文書一分割文書対応マップ生成部5003により生成された文書一分割文書対応マップを保存する文書一分割文書対応マップ保存部（文書一分割文書対応マップ保存手段）5009を備えている。なお、上記各保存部はたとえば共有のハードディスクや半導体メモリなどにより構成される。

【0343】

上記した構成により、本実施の形態の文書保存部5007は、文書データの内容や、文書の作成者、作成日、最終修正日などの文書データに付随する情報を適切な形式で保存する。また、文書データが文書内容とともにその要素からなる計量的な特徴ベクトルを持つ場合にはこれらも保存する。文書入力部5001にて、個々の文書データにそれらを一意に表す識別子が付与される場合にはこの識別子も適切な形式で保存することができる。

【0344】

また、分割文書保存部5008は、文書分割部5002により生成される分割文書データの内容を適切な形式で保存するとともに、計量的な特徴ベクトルを持つ場合にはこれらも保存する。個々の上記分割文書データにそれらを一意に表す識別子が付与される場合にはこの識別子も適切な形式で保存することができる。

【0345】

また、文書一分割文書対応マップ保存部5009は、文書一分割文書対応マップ生成部5003により生成される文書一分割文書対応マップを適切な形式で保存する。

【0346】

このように、実施の形態8によれば、文書データ、分割文書データ、および文書一分割文書対応マップが保存されるので、分割文書データおよび文書一分割文書対応マップを再生成することなしに、同一の文書データに対して、分類数、分類手法、または分類時の諸設定などパラメータの異なる分類結果を効率的に求めることができる。また、文書データを分類し、分類結果を生成するために必要なデータが保存されることにより、利用者は、分類作業に対して時間的な自由度を持つことができ、過去に行った文書分類の再分析を任意の時間におこなうこともできる。

【0347】

〔実施の形態9〕

図48は本発明の実施の形態9を示す文書分類装置の構成ブロック図である。図48に示したように、本実施の形態の文書分類装置は、図47に示した実施の形態8の構成に加え、分割文書分類結果生成部5005により生成された分割文書分類結果情報を保有する分割文書分類結果保存部（分割文書分類結果保存手段）5010を備えている。なお、上記分割文書分類結果保存部5010は、たとえば、共有のハードディスクや半導体メモリなどにより構成される。

【0348】

このように、第3の実施の形態によれば、文書データ、分割文書データ、文書一分割文書対応マップ、および、分割文書分類結果情報が保存されるので、実施の形態8の効果に加え、一度分類を実行すれば、その分類結果をテキスト表現や表表現やグラフ表現などさまざまな形式で表現することができる。また、分割文書分類結果情報が保存されることにより、分類の実行作業および分類結果の分析作業において、利用者は、時間的な自由度を持つことができ、過去に行った文書分類結果の再分析をさまざまな表現形式で任意の時間におこなうこともできる。

【0349】

【実施の形態10】

この発明の実施の形態10では、前記各実施の形態の文書分類装置、文書分類方法において、図49に示すように、文書分割部5002により生成される複数の分割文書データ中に分割前の文書データである文書1を含む。これにより、本実施の形態では、利用者は、分割されている文書データを分類することで得られる詳細な文書データの分類構造だけでなく、分割前の文書データ自体を分類した結果として得られるマクロな分類構造の融合した分類構造を得ることができる。

【0350】

【実施の形態11】

この発明の実施の形態11では、前記各実施の形態の文書分類装置、文書分類方法において、文書分割部5002は、文書データの構造情報を基に文書データを分割する。図50に、分類対象文書データがHTML形式で記述された文書の例を示す。分割をおこなう前に、図50に示したようなHTML形式の文書データから構造情報を抽出し、それらの構造を用いて文書の適切な分割規則を設定することにより文書データから分割文書データを生成する。

【0351】

つまり、この例では、文書データ中のタグ<L1>に着目し、「タグ<L1>を持つテキストを一つの分割文書データとする」という文言を分割文書データを生成する規則とする。この規則を文書データに適用することにより図50に示したような7つの分割文書が生成される。

【0352】

上記のように、文書が、HTML、XML、SGMLなど特定の構造化文書の形式を有していない場合でも、文字の大きさ、文字の装飾、文字の色、およびフォントなどに関する情報から分割規則を生成し、分割文書を生成することもできる。また、文書データがイメージであってOCR装置などにより入力される場合には、元のイメージのレイアウト情報などを利用することにより分割規則を生成し、分割文書を生成することもできる。

【0353】

なお、文書データのすべてをいずれかの分割文書データにする必要はない。た

とえば、図50に示した例では、文字列「ニューストピック（98/09/25）」は分割文書には採用しない。

【0354】

このように、実施の形態11では、文書データから構造情報を抽出し、文書割をおこなう前に構造情報を用いて文書の適切な分割規則を設定することにより、異なった話題の分割などを適切におこなうことができ、したがって、文書データの詳細な分類構造がわかる文書分類を適切におこなうことができる。

【0355】

〔実施の形態12〕

この発明の実施の形態12では、前記実施の形態7～10の文書分類装置、文書分類方法において、図51に示すように、文書データに含まれる単語など要素を抽出する文書要素解析部（文書要素抽出手段）5011、上記文書要素解析部5011により抽出された要素に付随する品詞など要素付隨情報を抽出する要素付隨情報抽出部（要素付隨情報抽出手段）5012を備え（図51は図48に示した実施の形態9に文書要素抽出部5011、要素付隨情報抽出5012を加えた例で示している）、文書分割部5002が、上記文書要素解析部5011により抽出された要素、または上記要素と上記要素付隨情報抽出部5012により抽出された要素付隨情報を用いて上記文書データを分割する。

【0356】

図52に示すように、文書分割をおこなう前に、自然言語処理手段である文書要素解析部5011が文書データから単語などそれらの要素を抽出し、要素付隨情報抽出部5012が品詞など要素付隨情報を抽出して文書の適切な分割規則を設定するのである。なお、上記文書要素解析部5011および要素付隨情報抽出部5012は新たに設けるのではなく、分割文書分類部5004内の同様の手段を用いることが可能である。

【0357】

この実施の形態では、たとえば、図52に示したように、文書データが特定の構造情報を持たない複数のニューストピックの集まりであり、各トピックが、単語「トピック」+「数字」+「改行コード」という文字列の後に記述されている

場合で説明すると、上記のような構造が文書要素解析部5011および要素付隨情報抽出部5012の抽出結果から認識され、文章の終端を考慮して、「トピック+数字+改行コードという文字列を先頭とし、上記文字列または文書終端記号を終端として囲まれる文字列を一つの分割文書データとする」という分割規則が生成されることになる。

【0358】

さらに詳しく説明すると、抽出された単語とその品詞情報などから、まず、名詞と改行コードのみを抽出し、つぎに、文字列「トピック+数字+改行コード」および文書終端記号を検出し、文書内でのそれらの位置を記憶する。そして、文書データに対して前記分割規則を適用し、図52に示したような分割文書データを生成する。

【0359】

なお、文書データのすべてをいずれかの分割文書データにする必要はなく、たとえば、図52に示した例では、文字列「ニューストピック（98/09/25）」は分割文書には採用しない。また、上記の例では、文書データから要素およびその付隨情報を抽出して分割規則を設定する場合で説明したが、要素のみを抽出してその要素情報から分割規則を設定することも可能である。

【0360】

こうして、実施の形態12によれば、文書データからそれらの要素情報などを抽出し、抽出した要素情報などを用いて文書の分割規則を設定することにより、実施の形態11と同様に、文書データの詳細な分類構造がわかる文書分類を適切におこなうことができる。

【0361】

【実施の形態13】

この発明の実施の形態13では、前記実施の形態7～10の文書分類装置、文書分類方法において、利用者により指示された指定範囲にしたがって文書分割部5002が文書データを分割する。図53に示すような文書データに対して利用者がそれぞれの分割文書の範囲を指定すると、指定にしたがって文書分割部5002が文書分割をおこなう。

【0362】

本実施の形態では、文書分割時、文書分割部5002がまず、画面上に、その初期状態として左右の指示ポイントおよび領域指定ラインからなる領域指定オブジェクトを文書の最上部に表示する。この状態で、利用者は、マウスなどポインティングデバイスを用いて、左右どちらかの指示ポイントをドラッグし、それを上下に移動させることにより、それぞれの分割文書の領域を選択することができる。

【0363】

また、この指定時、文書分割部5002は、領域選択処理をおこなっていることを示すため、指示ポインタを黒色から白色に、領域指定ラインを実線から破線に変化させる。選択領域を決定するには、所望の位置で指示ポイントのドラッグを止めればよい。

【0364】

つぎに、利用者は選択した領域を分割文書とするかしないか決定する。分割領域としない場合には、それを明示的に表示するために、文書分割部5002は選択領域を図示のように網掛け表示にさせる。

【0365】

こうして、本実施の形態によれば、利用者は文書データからそれぞれの分割文書データを所望通りに選択することができるので、文書データの詳細な分類構造がわかり、かつ利用者の意図に合った文書分類をおこなうことができる。

【0366】

【実施の形態14】

この発明の実施の形態14では、前記実施の形態7~10の文書分類装置、文書分類方法において、文書データ中の文字数、文数、または文字数と文数の両方を基に文書データを分割する。たとえば、図54に示す文書データをほぼ200文字を単位として分割をおこなう。

【0367】

ここで、ほぼ200文字を単位とするのは、正確な200文字単位としてもその終端が句点である保証がないことから、200文字目の前後のもっとも近い句

点をそれぞれの分割文書の終端とするからである。こうして、図54に示したような分割文書が生成される。同様に、所定の文数を単位とした文書分割をおこなうこともできるし、文字数と文数の両方を基にした文書分割をおこなうこともできる。

【0368】

このように、実施の形態14によれば、文字数、文数、または文字数と文数の両方を基に文書データを分割することにより、話題の異なった内容などが異なった分割文書として分割され、分類される可能性が高くなるので、文書データの詳細な分類構造がわかる文書分類をおこなうことができる。

【0369】

〔実施の形態15〕

この発明の実施の形態15では、前記各実施の形態の文書分類装置、文書分類方法において、文書分類結果生成部5006が分類結果情報として、文書データを示す情報および上記文書データに付随する代表的情報のみを提示する。

【0370】

たとえば図55に示すように、先頭に分類カテゴリ名を表示し、その横にそのカテゴリを代表するキーワードを表示し、カテゴリ名の下には文書データを示す情報として当該カテゴリに属する分割文書データを含んでいる文書データの、たとえば、文書データ名（文書名）を表示する。また、各文書データ名の左側には文書アイコンを表示させ、この文書アイコンが指示されたとき、文書データの内容を表示させる。

【0371】

また、各文書データ名の配置順は、カテゴリ代表値との類似度が高い分割文書データの文書データ名を先（左側）にする。また、同じ文書データから生成された複数の分割文書データが同一の分類カテゴリに属している場合には、類似度のもっとも高い分割文書データに対応する文書データ名のみを表示する。なお、上記キーワードとは出現頻度の多い単語である。

【0372】

このように、実施の形態15によれば、文書分類結果が文書データを示す情報

と文書データに付随する代表的情報のみが表示されるので、利用者は文書データの詳細な分類構造の概要を容易に把握することができる。

【0373】

【実施の形態16】

この発明の実施の形態16では、実施の形態15の文書分類結果提示に加えて、分割文書データを示す情報および上記分割文書データに付随する情報を提示する。

【0374】

たとえば、図56に示すように、先頭に分類カテゴリ名を表示し、その横にそのカテゴリを代表するキーワードを表示し、カテゴリ名の下には文書データを示す情報として当該カテゴリに属する分割文書データを含んでいる文書データのたとえば文書データ名（文書名）を表示する。

【0375】

また、各文書データ名の左側には文書アイコンを表示させ、この文書アイコンが指示されたとき、文書データの内容を表示させる。また、文書データ名の右側には分割文書アイコンを表示させる。なお、分割文書アイコン中には当該文書データにおける分割文書データの位置と当該文書データ中の分割文書数を表示させる。さらに、上記分割文書アイコンを指示することで文書データ中の当該分割文書データを表示させることができる。

【0376】

また、各文書データ名の配置順はカテゴリ代表値との類似度が高い分割文書データの文書データ名を先にする。また、同じ文書データから生成された複数の分割文書データが同一の分類カテゴリに属している場合には類似度の順位がわかるようにその順位を表示させる。

【0377】

このように、実施の形態16によれば、文書分類結果が文書データを示す情報と文書データに付随する代表的情報、および分割文書データを示す情報と分割文書データに付随する代表的情報のみが表示されるので、利用者は文書データの詳細な分類構造の概要とともにどの分割文書が起因して当該カテゴリに分類された

かというようなことも容易にわかる。

【0378】

以上、本発明の文書分類装置および文書分類方法を説明したが、この文書分類方法を実現するプログラムを着脱可能であるとともにコンピュータ読み取り可能な記録媒体に記録し、上記記録媒体を移した先の情報処理装置内で本発明によった文書分類をおこなうこともできる。

【0379】

【発明の効果】

以上説明したように、請求項1の発明によれば、入力された文書データを記憶する文書記憶手段と、前記文書記憶手段により記憶された文書データの全部または一部を選択する選択手段と、前記選択手段により選択された文書データの全部または一部から文字列の特徴に関するデータを抽出する特徴抽出手段と、前記特徴抽出手段により抽出された文字列の特徴に関するデータに基づいて前記文書データの全部または一部を加工処理する加工処理手段と、前記加工処理手段により加工処理された文書データの全部または一部を出力する出力手段とを備えるため、文書の意味に係わるような分析作業において、単にその結果のみを出力するのではなく、情報分析作業全般にわたる支援をおこなうことが可能な文書処理装置が得られるという効果を奏する。

【0380】

また、請求項2の発明によれば、前記出力手段が、前記加工処理手段により加工処理された文書データの全部または一部の内容に基づいて複数の項目値を設定する項目値設定手段と、前記項目値設定手段により設定された項目値ごとに前記文書データの全部または一部を集計する集計手段と、を備え、前記文書データの全部または一部を、項目値を少なくとも一つの軸とする表形式に展開して出力するため、簡易な操作で加工処理の結果をクロス表として表すことができ、情報の内容の把握を容易におこなうことができることから、文書の意味に係わるような分析作業において、単にその結果のみを出力するのではなく、情報分析作業全般にわたる支援をおこなうことが可能な文書処理装置が得られるという効果を奏する。

【0381】

また、請求項3の発明によれば、前記出力手段が、さらに、前記加工処理手段により加工処理された文書データの全部または一部を、前記加工処理手段により加工処理される前の文書データの全部または一部とともに出力するため、加工処理すべき対象データとその他のデータが同時に表示され、それを確認することにより、加工処理の対象範囲の決定を正確かつ容易におこなうことができることから、文書の意味に係わるような分析作業において、単にその結果のみを出力するのではなく、情報分析作業全般にわたる支援をおこなうことが可能な文書処理装置が得られるという効果を奏する。

【0382】

また、請求項4の発明によれば、前記文書記憶手段が、さらに、前記加工処理手段により加工処理された文書データの全部または一部を記憶するため、以後、他のデータと同様に扱うことができることから、文書の意味に係わるような分析作業において、単にその結果のみを出力するのではなく、情報分析作業全般にわたる支援をおこなうことが可能な文書処理装置が得られるという効果を奏する。

【0383】

また、請求項5の発明によれば、前記選択手段が、さらに、前記出力手段により出力された文書データの全部または一部を選択するため、出力手段により出力された文書データの全部または一部をさらなる分析の対象とすることができます、多彩で高度な情報分析作業ができることから、文書の意味に係わるような分析作業において、単にその結果のみを出力するのではなく、情報分析作業全般にわたる支援をおこなうことが可能な文書処理装置が得られるという効果を奏する。

【0384】

また、請求項6の発明によれば、前記文書記憶手段が、さらに、前記加工処理の内容に関するデータを記憶するため、加工処理の内容に関するデータの紛失を防止し、当該データの管理が容易になるだけでなく、加工処理に用いた設定とそれによる処理結果を関連づけて把握することができることから、文書の意味に係わるような分析作業において、単にその結果のみを出力するのではなく、情報分析作業全般にわたる支援をおこなうことが可能な文書処理装置が得られるという

効果を奏する。

【0385】

また、請求項7の発明によれば、入力手段が、文書データを入力し、言語解析手段が、前記入力手段により入力された文書データを解析して言語解析情報を得、ベクトル生成手段が、前記言語解析手段により得られた言語解析情報に基づいて前記文書データに対する文書特徴ベクトルを生成し、分類手段が、前記ベクトル生成手段により生成された文書特徴ベクトル間の類似度に基づいて文書を分類し、文書の部分集合を生成し、クラスタ特徴算出手段が、前記分類手段により生成された文書の部分集合の特徴であるクラスタ特徴を算出し、分類体系記憶手段が、前記クラスタ特徴算出手段により算出されたクラスタ特徴を分類体系の構成要素として記憶するため、クラスタを得ることができるとともに、クラスタ重心間の類似度等を用いて、クラスタの内容に基づくクラスタの構造化・体系化をおこなうことができ、これにより、任意の文書集合にどのような内容が含まれるかを漸次的に収集することが可能な文書分類装置が得られるという効果を奏する。

【0386】

また、請求項8の発明によれば、入力手段が、文書データを入力し、言語解析手段が、前記入力手段により入力された文書データを解析して言語解析情報を得、ベクトル生成手段が、前記言語解析手段により得られた言語解析情報に基づいて前記文書データに対する文書特徴ベクトルを生成し、分類手段が、前記ベクトル生成手段により生成された文書特徴ベクトル間の類似度に基づいて文書を分類し、文書の部分集合を生成し、クラスタ特徴算出手段が、前記分類手段により生成された文書の部分集合の特徴であるクラスタ特徴を算出し、表示手段が、前記クラスタ特徴算出手段により算出されたクラスタ特徴を表示し、クラスタ選択指示手段が、前記分類手段により生成された文書の部分集合の中から所望の部分集合を選択し、分類体系記憶手段が、前記クラスタ選択指示手段により選択された文書の部分集合を分類体系の構成要素として記憶するため、選択されたクラスタのみを用いて、より操作者の意図したものに近いクラスタの構造化・体系化をおこなうことができ、これにより、任意の文書集合にどのような内容が含まれるかを漸次的に収集することが可能な文書分類装置が得られるという効果を奏する。

【0387】

また、請求項9の発明によれば、請求項8の発明において、文書特徴ベクトル記憶手段が、前記ベクトル生成手段により生成された文書特徴ベクトルを記憶し、ベクトル修正手段が、前記文書特徴ベクトル記憶手段により記憶された文書特徴ベクトルを、前記クラスタ選択指示手段により選択された部分集合に属する文書の文書特徴ベクトルを除去したのこりとなるように修正し、前記分類手段が、前記ベクトル修正手段により修正された文書特徴ベクトルに基づいて文書を分類するため、既知になったクラスタの影響を排除した新たなクラスタを生成することができ、これにより、任意の文書集合にどのような内容が含まれるかを漸次的に収集することが可能な文書分類装置が得られるという効果を奏する。

【0388】

また、請求項10の発明によれば、請求項8の発明において、文書特徴ベクトル記憶手段が、前記ベクトル生成手段により生成された文書特徴ベクトルを記憶し、文書表現空間修正手段が、前記文書特徴ベクトル記憶手段により記憶された文書特徴ベクトル間の類似度を判断する際の文書表現空間を前記クラスタ選択指示手段により選択された部分集合から算出する特微量に基づいて修正し、前記分類手段が、前記文書表現空間修正手段により修正された文書表現空間を用いて、前記ベクトル生成手段により生成された文書特徴ベクトル間の類似度に基づいて文書を分類するため、前回の分類実行の結果、操作者に選択されたクラスタの形成特徴を次回の分類実行時に排除することができ、排除した状態で新たなクラスタを生成することができ、これにより、任意の文書集合にどのような内容が含まれるかを漸次的に収集することが可能な文書分類装置が得られるという効果を奏する。

【0389】

また、請求項11の発明によれば、請求項9の発明において、文書特徴ベクトル記憶手段が、前記ベクトル生成手段により生成された文書特徴ベクトルを記憶し、文書表現空間修正手段が、前記文書特徴ベクトル記憶手段により記憶された文書特徴ベクトル間の類似度を判断する際の文書表現空間を前記クラスタ選択指示手段により選択されたクラスタ特徴に基づいて修正し、前記分類手段が、前記

文書表現空間修正手段により修正された文書表現空間を用いて、前記ベクトル修正手段により修正された文書特徴ベクトル間の類似度に基づいて文書を分類するため、既知になったクラスタの影響を排除し、かつ、前回の分類実行の結果、操作者に選択されたクラスタの形成特徴を次回の分類実行時に排除することができ、排除した状態で新たなクラスタを生成することができ、これにより、任意の文書集合にどのような内容が含まれるかを漸次的に収集することができる文書分類装置が得られるという効果を奏する。

【0390】

また、請求項12の発明によれば、請求項8または10の発明において、選択情報付与手段が、前記分類手段により生成された文書の部分集合に所属する文書のすべてあるいは一部が選択された場合に選択されたことを示す選択情報を付与し、前記表示手段が、前記クラスタ特徴を表示するとともに、選択情報付与手段により付与された選択情報を表示するため、多重に利用される文書の識別性および一度も選択されない文書の識別性を向上させることができ、これにより、任意の文書集合にどのような内容が含まれるかを漸次的に収集することができる文書分類装置が得られるという効果を奏する。

【0391】

また、請求項13の発明によれば、請求項8～12の発明において、前記分類体系記憶手段が、前記選択指示手段により選択された文書の部分集合に属する全部あるいは一部の文書のほか、クラスタ特徴および／または操作者が作成した任意の情報を分類体系の構成要素として記憶するため、クラスタの内容把握を容易にし、かつ、操作者独自の分類体系を簡易に生成できるので、分類体系の利用価値を向上させることができ、これにより、任意の文書集合にどのような内容が含まれるかを漸次的に収集することができる文書分類装置が得られるという効果を奏する。

【0392】

また、請求項14の発明によれば、文書の内容にしたがって文書群を分類する文書分類装置において、文書データ群を入力する文書入力手段と、入力された文書データ群の各文書に対して所定の基準に基づき文書の分割をおこない、一つの

文書データから一つまたは複数の分割文書データを生成する文書分割手段と、前記文書データと前記分割文書データとの対応を示す文書一分割文書対応マップを生成する文書一分割文書対応マップ生成手段と、前記分割文書データを分類する分割文書分類手段と、前記分割文書分類手段による分類結果に基づいて分割文書分類結果情報を生成する分割文書分類結果生成手段と、前記文書一分割文書対応マップと前記分割文書分類結果情報を用いて前記文書データの分類結果情報を生成する文書分類結果生成手段と、を備えるため、一つの文書の中に複数の話題や意味が含まれている場合に、ある特定の話題や意味に限定されたカテゴリに分類されたり、利用者の意図するカテゴリとは異なるカテゴリに分類されたりすることがなく、したがって、利用者がその分類カテゴリをよく理解が可能で、また、分割前文書（所属文書）中の分割文書の位置なども示されるので、利用者は文書群中の読みみたい部分を効率的に読むことが可能な文書分類装置が得られるという効果を奏する。

【0393】

また、請求項15の発明によれば、請求項14の発明において、前記文書データを保存する文書保存手段と、前記分割文書データを保存する分割文書保存手段と、前記文書一分割文書対応マップ生成手段により生成された文書一分割文書対応マップを保存する文書一分割文書対応マップ保存手段と、を備えるため、分割文書データおよび文書一分割文書対応マップを再生成することなしに、同一の文書データに対して、分類数、分類手法、または分類時の諸設定などパラメータの異なる分類結果を効率的に求めることができ、また、文書データを分類し、分類結果を生成するために必要なデータが保存されることにより、利用者が分類作業に対して時間的な自由度を持つことが可能で、過去に行った文書分類の再分析を任意の時間間におこなうことも可能な文書分類装置が得られるという効果を奏する。

【0394】

また、請求項16の発明によれば、請求項15の発明において、前記分割文書分類結果生成手段により生成された分割文書分類結果情報を保存する分割文書分類結果保存手段を備えるため、請求項15の発明の効果に加え、一度分類を実行

すれば、その分類結果をテキスト表現や表表現やグラフ表現などさまざまな形式で表現することが可能で、また、分割文書分類結果情報が保存されることにより、分類の実行作業および分類結果の分析作業において、利用者が時間的な自由度を持つことが可能で、過去に行った文書分類結果の再分析をさまざまな表現形式で任意の時間におこなうことも可能な文書分類装置が得られるという効果を奏する。

【0395】

また、請求項17の発明によれば、請求項14～16の発明において、前記文書分割手段により生成される複数の分割文書データには分割前の文書データそのものを含むため、利用者は、分割されている文書データを分類することで得られる詳細な文書データの分類構造だけでなく、分割前の文書データ自体を分類した結果として得られる概略的でマクロな分類構造の融合した分類構造を得ることが可能な文書分類装置が得られるという効果を奏する。

【0396】

また、請求項18の発明によれば、請求項14～17の発明において、前記文書分割手段が、文書データの構造情報を基に文書データを分割する構成にしたため、異なった話題の分割等を適切におこなうことができ、したがって、文書データの詳細な分類構造がわかる文書分類を適切におこなうことが可能な文書分類装置が得られるという効果を奏する。

【0397】

また、請求項19の発明によれば、請求項14～17の発明において、前記文書データに含まれる要素を抽出する文書要素抽出手段と、前記文書要素抽出手段により抽出された要素に付随する要素付隨情報を抽出する要素付隨情報抽出手段と、を備え、前記文書分割手段が、前記文書要素抽出手段により抽出された要素、または前記要素と前記要素付隨情報抽出手段により抽出された要素付隨情報とを用いて前記文書データを分割する構成にしたため、文書データの詳細な分類構造がわかる文書分類を適切におこなうことが可能な文書分類装置が得られるという効果を奏する。

【0398】

また、請求項20の発明によれば、請求項14～17の発明において、前記文書分割手段が、指示された指定範囲にしたがって文書データの分割をおこなう構成にしたため、利用者の意図に合い、かつ文書データの詳細な分類構造がわかる文書分類をおこなうことが可能な文書分類装置が得られるという効果を奏する。

【0399】

また、請求項21の発明によれば、請求項14～17において、前記文書分割手段が、文書データ中の文字数、文数、または文字数と文数の両方を基に文書データを分割する構成にしたため、話題の異なった内容などが異なった文書として分類される可能性が高くなり、したがって、この発明でも文書データの詳細な分類構造がわかる文書分類をおこなうことが可能な文書分類装置が得られるという効果を奏する。

【0400】

また、請求項22の発明によれば、請求項14～21の発明において、前記文書分類結果生成手段が、文書データを示す情報および前記文書データに付随する代表的情報を、分類結果情報として抽出して提示する構成にしたため、利用者は文書データの詳細な分類構造の概要や全体的な構造を容易に把握することが可能な文書分類装置が得られるという効果を奏する。

【0401】

また、請求項23の発明によれば、請求項22の発明において、前記文書分類結果生成手段が、分割文書データを示す情報および前記分割文書データに付隨する代表的情報を、分類結果情報として、抽出して提示する構成にしたため、利用者は文書データの詳細な分類構造の概要や全体的な構造とともにどの分割文書が起因して当該カテゴリに分類されたかというようなことも容易にわかる文書分類装置が得られるという効果を奏する。

【0402】

また、請求項24の発明によれば、入力された文書データを記憶する文書記憶工程と、前記文書記憶工程により記憶された文書データの全部または一部を選択する選択工程と、前記選択工程により選択された文書データの全部または一部から文字列の特徴に関するデータを抽出する特徴抽出工程と、前記特徴抽出工程に

より抽出された文字列の特徴に関するデータに基づいて前記文書データの全部または一部を加工処理する加工処理工程と、前記加工処理工程により加工処理された文書データの全部または一部を出力する出力工程と、を含むので、文書の意味に係わるような分析作業において、単にその結果のみを出力するのではなく、情報分析作業全般にわたる支援をおこなうことが可能な文書処理方法が得られるという効果を奏する。

【0403】

また、請求項25の発明によれば、前記出力工程が、前記加工処理工程により加工処理された文書データの全部または一部の内容に基づいて複数の項目値を設定する項目値設定工程と、前記項目値設定工程により設定された項目値ごとに前記文書データの全部または一部を集計する集計工程と、を含み、前記文書データの全部または一部を、項目値を少なくとも一つの軸とする表形式に展開して出力するので、簡易な操作で加工処理の結果をクロス表として表すことができ、情報の内容の把握を容易におこなうことができることから、文書の意味に係わるような分析作業において、単にその結果のみを出力するのではなく、情報分析作業全般にわたる支援をおこなうことが可能な文書処理方法が得られるという効果を奏する。

【0404】

また、請求項26の発明によれば、前記出力工程が、さらに、前記加工処理工程により加工処理された文書データの全部または一部を、前記加工処理工程により加工処理される前の文書データの全部または一部とともに出力するので、加工処理すべき対象データとその他のデータが同時に表示され、それを確認することにより、加工処理の対象範囲の決定を正確かつ容易におこなうことができることから、文書の意味に係わるような分析作業において、単にその結果のみを出力するのではなく、情報分析作業全般にわたる支援をおこなうことが可能な文書処理方法が得られるという効果を奏する。

【0405】

また、請求項27の発明によれば、前記文書記憶工程が、さらに、前記加工処理工程により加工処理された文書データの全部または一部を記憶するので、以後

、他のデータと同様に扱うことができることから、文書の意味に係わるような分析作業において、単にその結果のみを出力するのではなく、情報分析作業全般にわたる支援をおこなうことが可能な文書処理方法が得られるという効果を奏する。

【0406】

また、請求項28の発明によれば、前記選択工程が、さらに、前記出力工程により出力された文書データの全部または一部を選択するので、出力工程により出力された文書データの全部または一部をさらなる分析の対象とすることができ、多彩で高度な情報分析作業ができることから、文書の意味に係わるような分析作業において、単にその結果のみを出力するのではなく、情報分析作業全般にわたる支援をおこなうことが可能な文書処理方法が得られるという効果を奏する。

【0407】

また、請求項29の発明によれば、前記文書記憶工程が、さらに、前記加工処理の内容に関するデータを記憶するので、加工処理の内容に関するデータの紛失を防止し、当該データの管理が容易になるだけでなく、加工処理に用いた設定とそれによる処理結果を関連づけて把握することができることから、文書の意味に係わるような分析作業において、単にその結果のみを出力するのではなく、情報分析作業全般にわたる支援をおこなうことが可能な文書処理方法が得られるという効果を奏する。

【0408】

また、請求項30の発明によれば、入力工程が、文書データを入力し、言語解析工程が、前記入力工程により入力された文書データを解析して言語解析情報を得、ベクトル生成工程が、前記言語解析工程により得られた言語解析情報に基づいて前記文書データに対する文書特徴ベクトルを生成し、分類工程が、前記ベクトル生成工程により生成された文書特徴ベクトル間の類似度に基づいて文書を分類し、文書の部分集合を生成し、クラスタ特徴算出工程が、前記分類工程により生成された文書の部分集合の特徴であるクラスタ特徴を算出し、分類体系生成工程が、前記クラスタ特徴算出工程により算出されたクラスタ特徴に基づいて分類体系の構成要素を生成するので、クラスタを得ることができるとともに、クラス

タ重心間の類似度等を用いて、クラスタの内容に基づくクラスタの構造化・体系化をおこなうことができ、これにより、任意の文書集合にどのような内容が含まれるかを漸次的に収集することが可能な文書分類方法が得られるという効果を奏する。

【0409】

また、請求項31の発明によれば、入力工程が、文書データを入力し、言語解析工程が、前記入力工程により入力された文書データを解析して言語解析情報を得、ベクトル生成工程が、前記言語解析工程により得られた言語解析情報に基づいて前記文書データに対する文書特徴ベクトルを生成し、分類工程が、前記ベクトル生成工程により生成された文書特徴ベクトル間の類似度に基づいて文書を分類し、文書の部分集合を生成し、クラスタ特徴算出工程が、前記分類工程により生成された文書の部分集合の特徴であるクラスタ特徴を算出し、表示工程が、前記クラスタ特徴算出工程により算出されたクラスタ特徴を表示し、クラスタ選択指示工程が、前記分類工程により生成された文書の部分集合の中から所望の部分集合を選択し、分類体系生成工程が、前記クラスタ選択指示工程により選択されたクラスタ特徴に基づいて分類体系の構成要素を生成するので、選択されたクラスタのみを用いて、より操作者の意図したものに近いクラスタの構造化・体系化をおこなうことができ、これにより、任意の文書集合にどのような内容が含まれるかを漸次的に収集することが可能な文書分類方法が得られるという効果を奏する。

【0410】

また、請求項32の発明によれば、請求項31の発明において、ベクトル修正工程が、前記クラスタ選択指示手段により選択された部分集合に属する文書の文書特徴ベクトルを除去したのこりとなるように修正し、前記分類工程が、前記ベクトル修正工程により修正された文書特徴ベクトルに基づいて文書を分類するので、既知になったクラスタの影響を排除した新たなクラスタを生成することができ、これにより、任意の文書集合にどのような内容が含まれるかを漸次的に収集することが可能な文書分類方法が得られるという効果を奏する。

【0411】

また、請求項33の発明によれば、請求項31の発明において、文書表現空間修正工程が、前記ベクトル生成工程により生成された文書特徴ベクトル間の類似度を判断する際の文書表現空間を前記クラスタ選択指示工程により選択された部分集合から算出する特徴量に基づいて修正し、前記分類工程が、前記文書表現空間修正工程により修正された文書表現空間を用いて、前記ベクトル生成手段工程により生成された文書特徴ベクトル間の類似度に基づいて文書を分類するので、前回の分類実行の結果、操作者に選択されたクラスタの形成特徴を次回の分類実行時に排除することができ、排除した状態で新たなクラスタを生成することができ、これにより、任意の文書集合にどのような内容が含まれるかを漸次的に収集することが可能な文書分類方法が得られるという効果を奏する。

【0412】

また、請求項34の発明によれば、請求項32の発明において、文書表現空間修正工程が、前記ベクトル生成工程により生成された文書特徴ベクトル間の類似度を判断する際の文書表現空間を前記クラスタ選択指示工程により選択された部分集合から算出する特徴量に基づいて修正し、前記分類工程が、前記文書表現空間修正工程により修正された文書表現空間を用いて、前記ベクトル修正工程により修正された文書特徴ベクトル間の類似度に基づいて文書を分類するので、既知になったクラスタの影響を排除し、かつ、前回の分類実行の結果、操作者に選択されたクラスタの形成特徴を次回の分類実行時に排除することができ、排除した状態で新たなクラスタを生成することができ、これにより、任意の文書集合にどのような内容が含まれるかを漸次的に収集することが可能な文書分類方法が得られるという効果を奏する。

【0413】

また、請求項35の発明によれば、請求項31または33の発明において、選択情報付与工程が、前記分類工程により生成された文書の部分集合に所属する文書のすべてあるいは一部が選択された場合に選択されたことを示す選択情報を付与し、前記表示工程が、前記クラスタ特徴を表示するとともに、選択情報付与工程により付与された選択情報を表示するので、多重に利用される文書の識別性および一度も選択されない文書の識別性を向上させることができ、これにより、任

意の文書集合にどのような内容が含まれるかを漸次的に収集することが可能な文書分類方法が得られるという効果を奏する。

【0414】

また、請求項36の発明によれば、請求項31～35の発明において、前記分類体系生成工程が、前記選択指示工程により選択されたクラスタ特徴のほか、前記文書の部分集合の中から選択された文書の部分集合に所属する文書群の全部あるいは一部および／または操作者が作成した情報に基づいて分類体系の構成要素を生成するので、クラスタの内容把握を容易にし、かつ、操作者独自の分類体系を簡易に生成できることで、分類体系の利用価値を向上させることができ、これにより、任意の文書集合にどのような内容が含まれるかを漸次的に収集することが可能な文書分類方法が得られるという効果を奏する。

【0415】

また、請求項37の発明によれば、一つの文書の中に複数の話題や意味が含まれている場合に、ある特定の話題や意味に限定されたカテゴリに分類されたり、利用者の意図するカテゴリとは異なるカテゴリに分類されたりすることなく、したがって、利用者がその分類カテゴリをよく理解できる。また、分割前文書（所属文書）中の分割文書の位置なども示されるので、利用者は文書群中の読みたい部分を効率的に読むことが可能な文書分類方法が得られるという効果を奏する。

【0416】

また、請求項38の発明によれば、請求項24～37のいずれか一つに記載された方法をコンピュータに実行させるプログラムを記録したことで、そのプログラムを機械読み取り可能となり、これによって、請求項24～37の動作をコンピュータによって実現することが可能な記録媒体が得られるという効果を奏する。

【図面の簡単な説明】

【図1】

この発明の実施の形態1による文書処理装置を構成する情報処理システム全体のハードウェア構成を示す説明図である。

【図2】

この発明の実施の形態1による文書処理装置を構成する情報処理システムにおけるサーバーのハードウェア構成を示す説明図である。

【図3】

この発明の実施の形態1による文書処理装置を構成する情報処理システムにおけるクライアントのハードウェア構成を示す説明図である。

【図4】

この発明の実施の形態1による文書処理装置の構成を機能的に示すブロック図である。

【図5】

この発明の実施の形態1による文書処理装置の項目名と項目値の関係を示す説明図である。

【図6】

この発明の実施の形態1による文書処理装置の文書記憶部に記憶された文書のデータ構造を示す説明図である。

【図7】

この発明の実施の形態1による文書処理装置の文書記憶部に記憶された文書の別のデータ構造を示す説明図である。

【図8】

この発明の実施の形態1による文書処理装置の出力部による画面表示の例を示す説明図である。

【図9】

この発明の実施の形態1による文書処理装置の出力部による画面表示の別の例を示す説明図である。

【図10】

この発明の実施の形態1による文書処理装置の出力部による画面表示の別の例を示す説明図である。

【図11】

この発明の実施の形態1による文書処理装置の特徴抽出部によりおこなわれる

抽出処理の内容の一覧を示す説明図である。

【図12】

この発明の実施の形態1による文書処理装置の加工処理部によりおこなわれる加工処理の内容の一覧を示す説明図である。

【図13】

この発明の実施の形態1による文書処理装置の各項目の特徴ベクトルを示す説明図である。

【図14】

この発明の実施の形態1による文書処理装置の単語とその単語IDごとの出現回数を示す説明図である。

【図15】

この発明の実施の形態1による文書処理装置の出力部による画面表示の別の例を示す説明図である。

【図16】

この発明の実施の形態1による文書処理装置の出力部によるクロス表作成のための指示画面を示す説明図である。

【図17】

この発明の実施の形態1による文書処理装置の出力部による分類処理の結果が表示されたクロス表を示す説明図である。

【図18】

この発明の実施の形態1による文書処理装置の出力部による分類処理の結果が表示された別のクロス表を示す説明図である。

【図19】

この発明の実施の形態1による文書処理装置の出力部の詳細な構成を示すブロック図である。

【図20】

この発明の実施の形態1による文書処理装置のクロス表の出力手順を示すフローチャートである。

【図21】

この発明の実施の形態1による文書処理装置の出力部による画面表示の別の例を示す説明図である。

【図22】

この発明の実施の形態1による文書処理装置の出力部による画面表示の別の例を示す説明図である。

【図23】

この発明の実施の形態1による文書処理装置の出力部による画面表示の別の例を示す説明図である。

【図24】

この発明の実施の形態1による文書処理装置の出力部による画面表示の別の例を示す説明図である。

【図25】

この発明の実施の形態1による文書処理装置の文書記憶部の詳細な構成を示すブロック図である。

【図26】

この発明の実施の形態1による文書処理装置の出力部による画面表示の別の例を示す説明図である。

【図27】

この発明の実施の形態1による文書処理装置の出力部による画面表示の別の例を示す説明図である。

【図28】

この発明の実施の形態1による文書処理装置の出力部による画面表示の別の例を示す説明図である。

【図29】

この発明の実施の形態1による文書処理装置の文書処理の一連の手順を示すフローチャートである。

【図30】

この発明の実施の形態2による文書分類装置の構成を機能的に示すブロック図である。

【図31】

この発明の実施の形態2による文書分類装置のクラスタ特徴表示部の表示の一例を示す説明図である。

【図32】

この発明の実施の形態2による文書分類装置の一連の処理の手順を示すフローチャートである。

【図33】

この発明の実施の形態3による文書分類装置の構成を機能的に示すブロック図である。

【図34】

この発明の実施の形態3による文書分類装置の一連の処理の手順を示すフローチャートである。

【図35】

この発明の実施の形態4による文書分類装置の構成を機能的に示すブロック図である。

【図36】

この発明の実施の形態4による文書分類装置の一連の処理の手順を示すフローチャートである。

【図37】

この発明の実施の形態5による文書分類装置の構成を機能的に示すブロック図である。

【図38】

この発明の実施の形態5による文書分類装置の一連の処理の手順を示すフローチャートである。

【図39】

この発明の実施の形態6による文書分類装置の構成を機能的に示すブロック図である。

【図40】

この発明の実施の形態6による文書分類装置の分類結果記憶部において設けら

れたテーブルを示す説明図である。

【図4 1】

この発明の実施の形態6による文書分類装置の選択情報付与部の処理の手順を示すフローチャートである。

【図4 2】

この発明の実施の形態7を示す文書分類装置の構成ブロック図である。

【図4 3】

この発明の実施の形態7による文書分類装置および文書分類方法の説明図である。

【図4 4】

この発明の実施の形態7による文書分類装置および文書分類方法の他の説明図である。

【図4 5】

この発明の実施の形態7による文書分類装置および文書分類方法の他の説明図である。

【図4 6】

この発明の実施の形態7による文書分類装置および文書分類方法の他の説明図である。

【図4 7】

この発明の実施の形態8による文書分類装置の構成ブロック図である。

【図4 8】

この発明の実施の形態9による文書分類装置の構成ブロック図である。

【図4 9】

この発明の実施の形態10による文書分類装置および文書分類方法の説明図である。

【図5 0】

この発明の実施の形態11による文書分類装置および文書分類方法の説明図である。

【図5 1】

この発明の実施の形態12による文書分類装置の構成ブロック図である。

【図52】

この発明の実施の形態12による文書分類装置および文書分類方法の説明図である。

【図53】

この発明の実施の形態13による文書分類装置および文書分類方法の説明図である。

【図54】

この発明の実施の形態14による文書分類装置および文書分類方法の説明図である。

【図55】

この発明の実施の形態15による文書分類装置および文書分類方法の説明図である。

【図56】

この発明の実施の形態16による文書分類装置および文書分類方法の説明図である。

【符号の説明】

101 サーバー

102 クライアント

103 ネットワーク

201 CPU

204 I/F

206 ディスク装置

301 CPU

306 ハードディスク

308 ディスプレイ

309 I/F

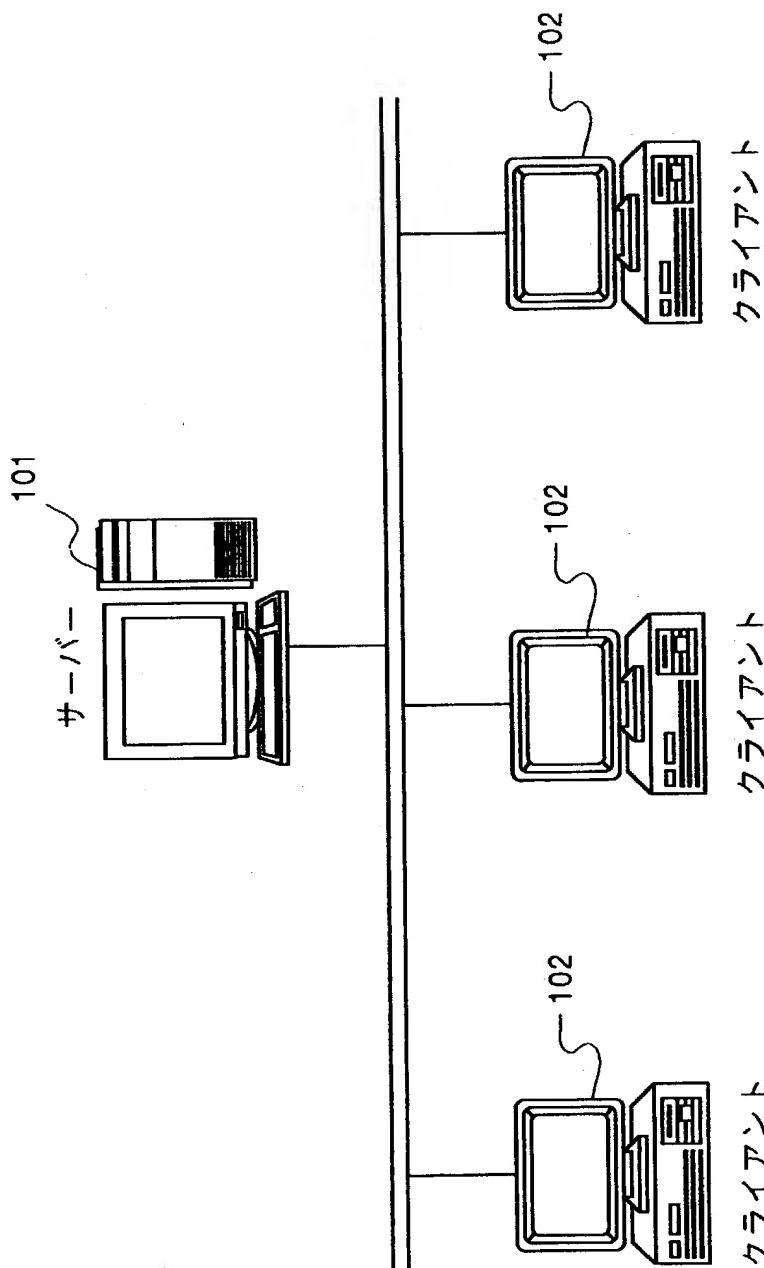
311 キーボード

312 マウス

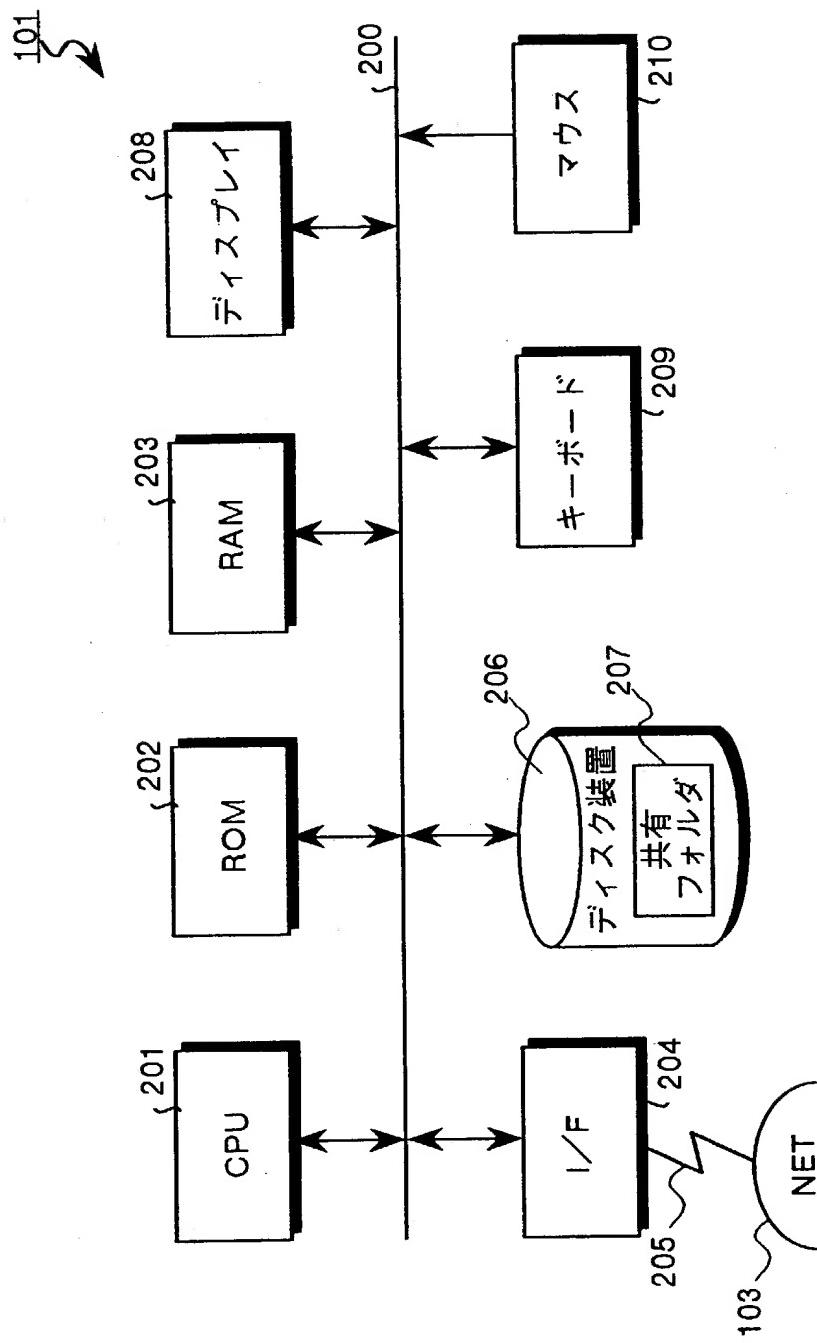
313 スキャナ
400 バス
401 入力部
402 文書記憶部
403 選択部
404 特徴抽出部
405 加工処理部
406 出力部
407 グラフ描画部
408 加工処理結果保持部
409 解析部
410 特徴ベクトル生成部
1600 指示画面
1700 クロス表
1800 マウスポインタ
1801 内容表示画面
1901 項目値選定部
1902 集計部
1903 表保持部
2401 棒グラフ表示領域
2501 設定値記憶部
2502 設定値送受信部
2503 分類情報記憶部
2604 問い合わせ画面
2703 分類情報表示画面
2802 表示領域
3001 入力部
3002 言語解析部
3003 ベクトル生成部

- 3004 分類部
- 3005 分類パラメータ指示部
- 3006 分類結果記憶部
- 3007 クラスタ特徴表示部
- 3008 クラスタ特徴算出部
- 3009 分類体系記憶部
- 3010 クラスタ選択指示部
- 3011 分類体系閲覧操作部
- 3110 カーソル
- 3301, 3501, 3701 ベクトル記憶部
- 3302, 3702 ベクトル修正部
- 3502, 3703 文書表現空間修正部
- 3901 選択情報付与部
- 4000 テーブル
- 5001 文書入力部
- 5002 文書分割部
- 5003 文書一分割文書対応マップ生成部
- 5004 分割文書分類部
- 5005 分割文書分類結果生成部
- 5006 文書分類結果生成部
- 5007 文書保存部
- 5008 分割文書保存部
- 5009 文書一分割文書対応マップ保存部
- 5010 分割文書分類結果保存部
- 5011 文書要素解析部
- 5012 要素付隨情報抽出部

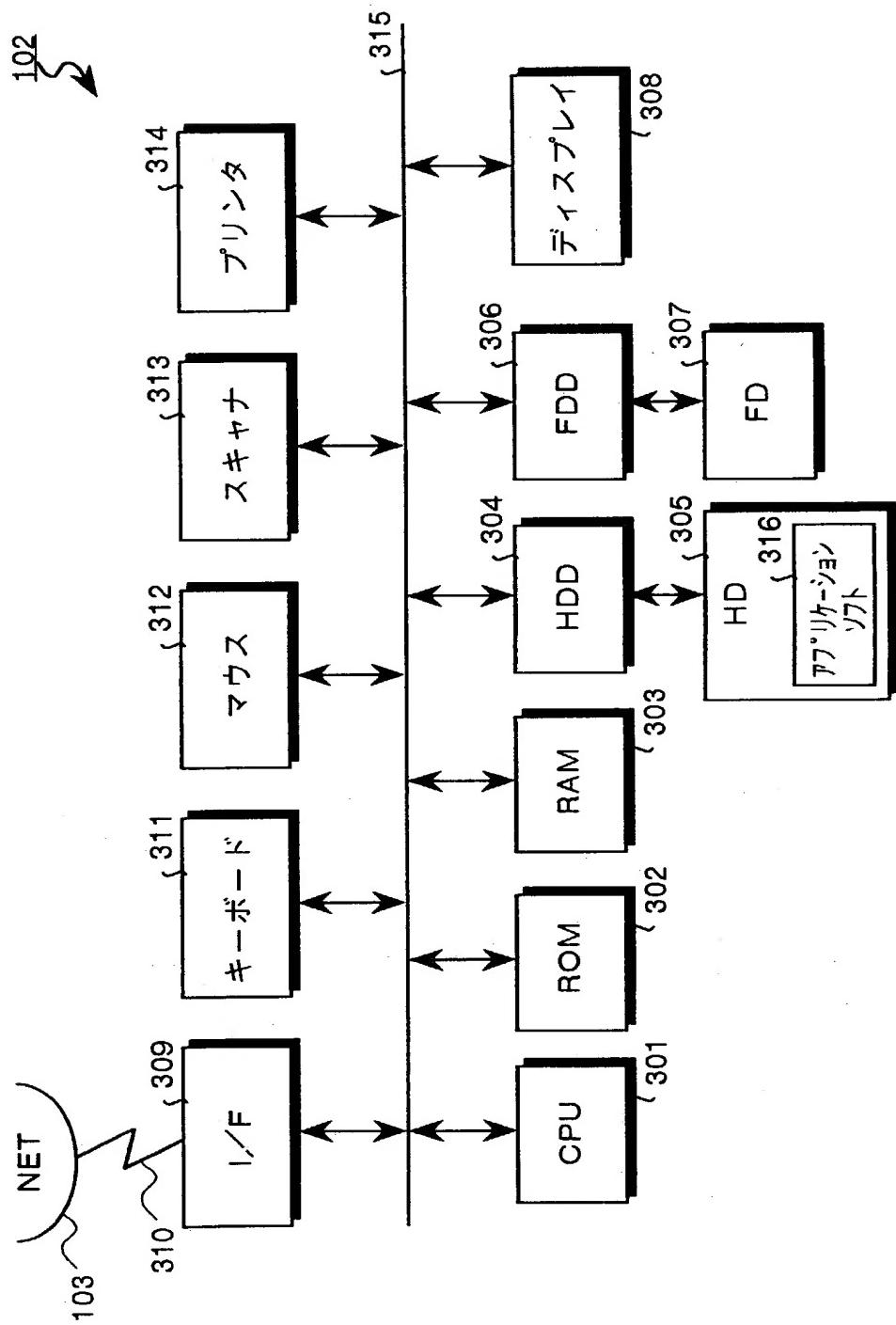
【書類名】 図面
【図1】



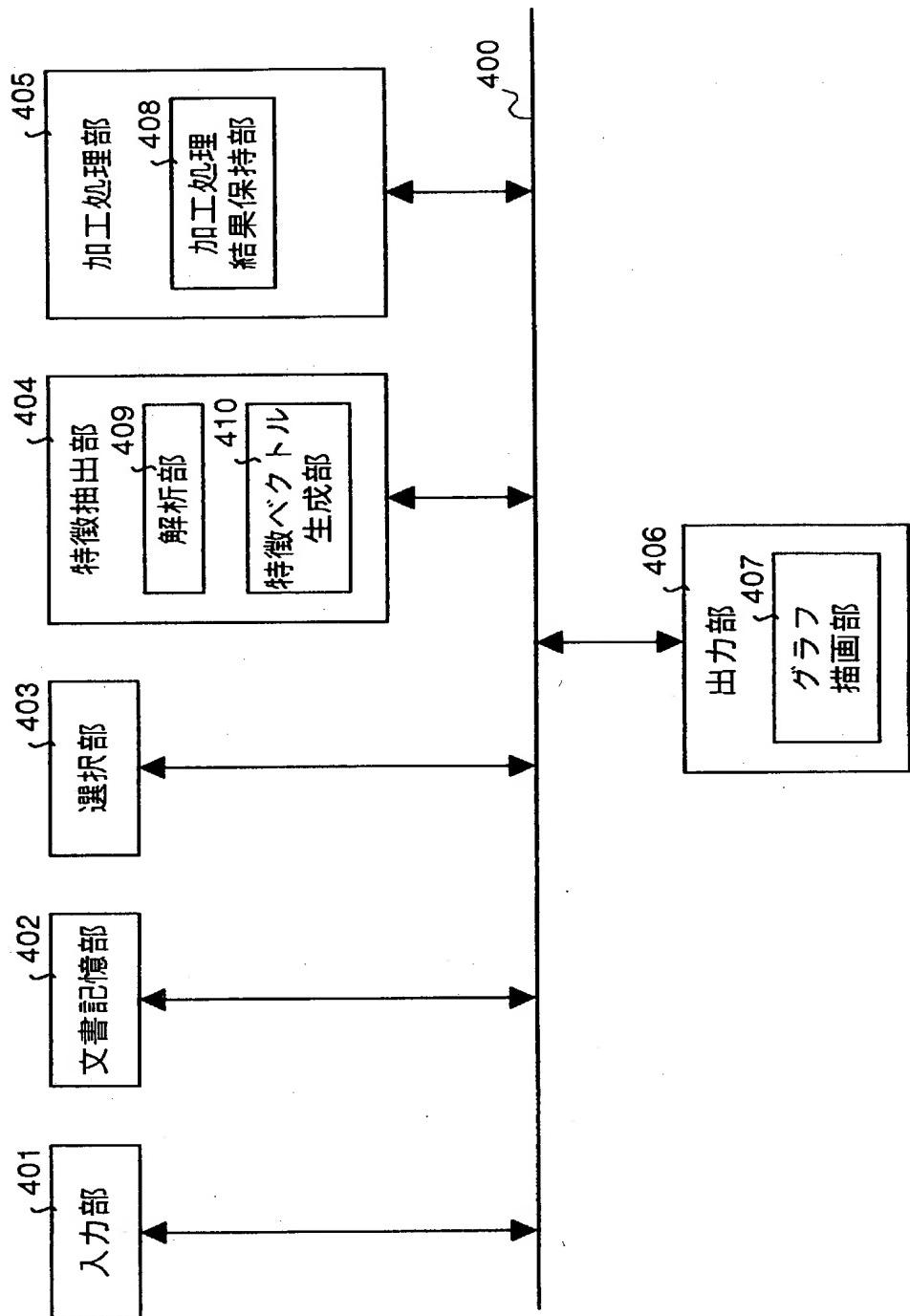
【図2】



【図3】



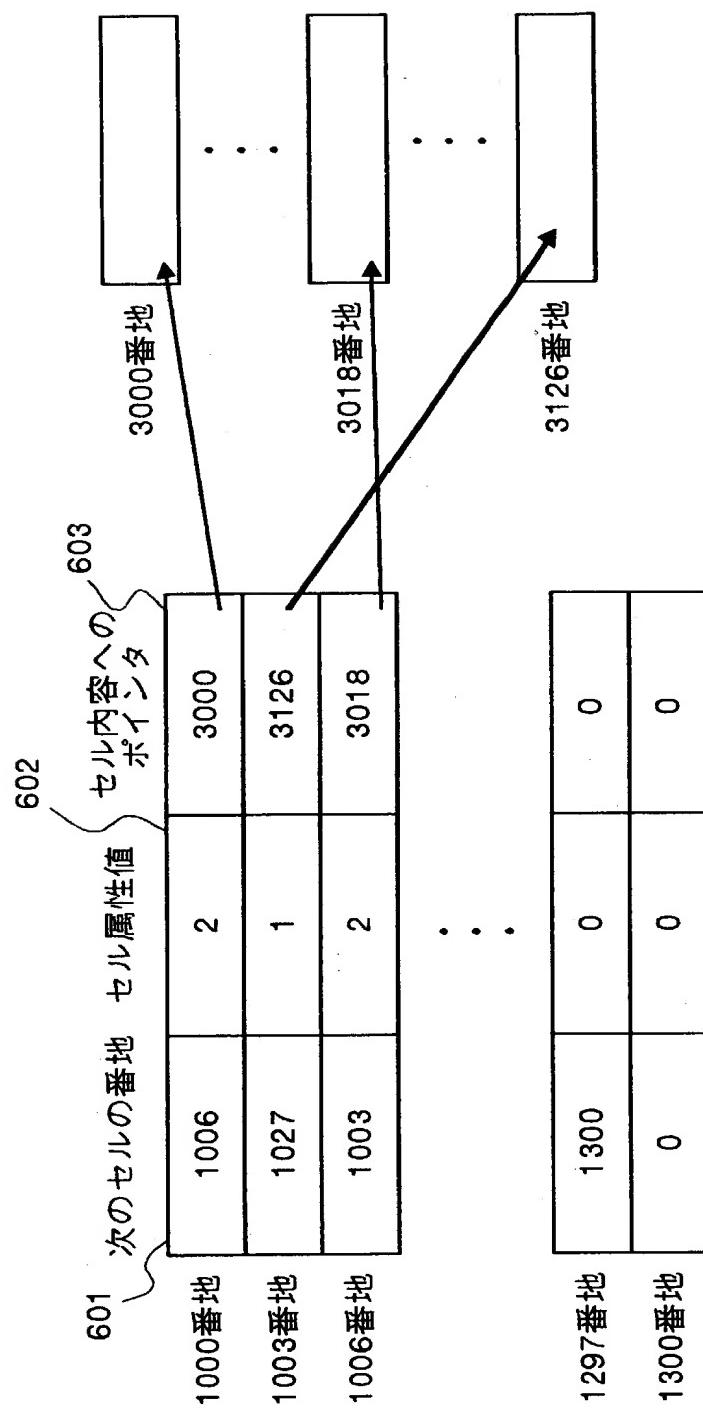
【図4】



【図5】

項目名	項目値
出願番号	特願平10-00000
出願日	平成10年1月1日
請求項の数	3
発明の名称	画像形成装置
発明の効果	本発明の画像形成装置は、○○○を××しているので…

【図6】



【図7】

	セル属性値			セル内容への ポイントタ			
701 1000番地	2	3000	2	3018	..	2	3327
1100番地	1	3421	1	3512	..	1	3698
..
1500番地	0	0	0	0	..	0	0

【図8】

番号	受付日	業所	車種	年式	内容
1	1997/3/5	愛知	ABC1800	1993	騒音が大きい
2	1997/3/5	富山	ABC2000	1995	排気が黒い
3	1997/3/5	京幌	ABC1800	1996	塗装が変色する
4	1997/3/5	福岡	DEF1600	1995	オイルが漏れる
5	1997/3/5	札幌	KLM1200	1992	暖房が効かない
6	1997/3/5	登別	DEF1600	1994	騒音が大きい
7	1997/3/5	長野	DEF1600	1996	エンジンがかからない
8	1997/3/5	東京	ABC1800	1997	オーバーヒートが起る
9	1997/3/6	松高	XYZ3000	1992	バッテリーが上がる
10	1997/3/6	崎長	KLM1200	1993	エンジンがかからない
11	1997/3/6	大阪	ABC1600	1994	排気が黒い
12	1997/3/6	野長	DEF1600	1997	ラジオが鳴らない
13	1997/3/6	岡盛	ABC1800	1996	塗装がはげる
14	1997/3/6	仙台	XYZ3000	1995	暖房が効かない

【図9】

番号	受付日	営業所	車種	年式	内容
11	1997/3/6	大阪	ABC1600	1994	排気が黒い
53	1997/4/21	長野	ABC1600	1993	オイルが漏れる
1	1997/3/5	愛知	ABC1800	1993	騒音が大きくなる
3	1997/3/5	東京	ABC1800	1996	オイルが漏れる
8	1997/3/5	東岡山	ABC1800	1997	騒音が大きい
13	1997/3/6	大阪	ABC1800	1996	オイル漏れ
18	1997/3/10	島	ABC1800	1994	オイル漏れ
28	1997/3/12	横浜	ABC1800	1995	騒音が大きい
35	1997/3/17	豊田	ABC1800	1996	オイル漏れ
39	1997/3/20	知	ABC1800	1993	騒音が大きい
42	1997/3/22	横浜	ABC1800	1996	オイル漏れ
46	1997/3/24	東京	ABC1800	1997	騒音が大きい
47	1997/3/24	富士	ABC1800	1994	オイル漏れ
4	1997/3/25	大阪	DEF1600	1995	オイル漏れ
901					

【図10】

1001

番号	受付日	業者所	車種	年式	内容
11	1997/3/6	大阪	ABC1600	1994	弁氣ガルが漏れる
53	1997/4/21	長野	ABC1600	1993	オ才イ音漏れ大気いする
1	1997/3/5	愛知	ABC1800	1993	騒音漏れがトドか起さる
3	1997/3/5	東京	ABC1800	1996	騒音漏れがトドか起さる
8	1997/3/5	岡山	ABC1800	1997	ハニタテ黒い
13	1997/3/6	大阪	ABC1800	1996	ハニタテ黒い
18	1997/3/10	島	ABC1800	1994	ハニタテ黒い
28	1997/3/12	横浜	ABC1800	1995	ハニタテ黒い
35	1997/3/17	知	ABC1800	1996	ハニタテ黒い
39	1997/3/20	東京	ABC1800	1993	ハニタテ黒い
42	1997/3/22	富山	ABC1800	1996	ハニタテ黒い
46	1997/3/24	大阪	ABC1800	1997	ハニタテ黒い
47	1997/3/24	札幌	ABC1800	1994	ハニタテ黒い
4	1997/3/25		DEF1600	1995	オイル漏れる

【図11】

	抽出処理内容
1	対象とする文字列に含まれる単語
2	対象とする文字列に含まれる単語数
3	対象とする文字列に含まれる単語の文字数
4	対象とする文字列に含まれる単語それぞれの出現回数
5	対象とする文字列に含まれる単語それぞれの品詞
6	対象とする文字列に含まれる単語間の関係の情報
7	対象とする文字列に含まれる文の数
8	対象とする文字列に含まれる文の文字数
9	対象とする文字列に含まれる文の文節数
10	対象とする文字列に含まれる文の間の関係
:	:

【図12】

加工処理内容	
1	分類処理
2	検索処理
3	並べ替え処理
4	代表値抽出処理
5	最大値抽出処理
6	最小値抽出処理
7	算術処理
:	:

【図13】

	騒音	が	大きい	塗装	変色	する	オーバーヒート	起こる	はげる	バッテリー	上がる	排気	黒い
騒音が大きい	1	1	1	0	0	0	0	0	0	0	0	0	0
塗装が変色する	0	1	0	1	1	0	0	0	0	0	0	0	0
オーバーヒートが起こる	0	1	0	0	0	0	1	1	0	0	0	0	0
塗装がはげる	0	1	0	1	0	0	0	0	1	0	0	0	0
バッテリーが上がる	0	1	0	0	0	0	0	0	0	0	1	1	0
排気が黒い	0	1	0	0	0	0	0	0	0	0	0	1	1

【図14】

単語	単語ID
騒音	1
が	2
大きい	3
塗装	4
変色	5
する	6
ヒート	7
起ころる	8
はげる	9
パーティ	10
上がる	11
排気	12
黒い	13

【図15】

番号	受付日	営業所	車種	年式	内容	クラクタ番号
1	1997/3/5	愛知	ABC1800	1993	騒音が大きい	5
2	1997/3/5	東京	ABC2000	1995	排気が黒い	1
3	1997/3/5	札幌	ABC1800	1996	塗装が変色する	7
4	1997/3/5	福岡	DEF1600	1995	オイルが漏れる	11
5	1997/3/5	別野	KLM1200	1992	暖房が効かない	2
6	1997/3/5	長野	DEF1600	1994	騒音が大きい	5
7	1997/3/5	東京	DEF1600	1996	エンジンがかからない	8
8	1997/3/5	高松	ABC1800	1997	オーバーヒートが起る	14
9	1997/3/6	長崎	XYZ3000	1992	バッテリーが上がる	12
10	1997/3/6	大阪	KLM1200	1993	エンジンがかからない	8
11	1997/3/6	長野	ABC1600	1994	排気が黒い	1
12	1997/3/6	盛岡	DEF1600	1997	ラジオが鳴らない	6
13	1997/3/6	仙台	ABC1800	1996	塗装がはげる	7
14	1997/3/6		XYZ3000	1995	暖房が効かない	2

【図16】

1600

<クロス表の作成		
処理対象項目名 :	(1601 内容 []	
軸となる項目名 :	(1602 車種 []	縦軸 横軸 []
分類数 :	(1605 [] 50	(1603 1604 []

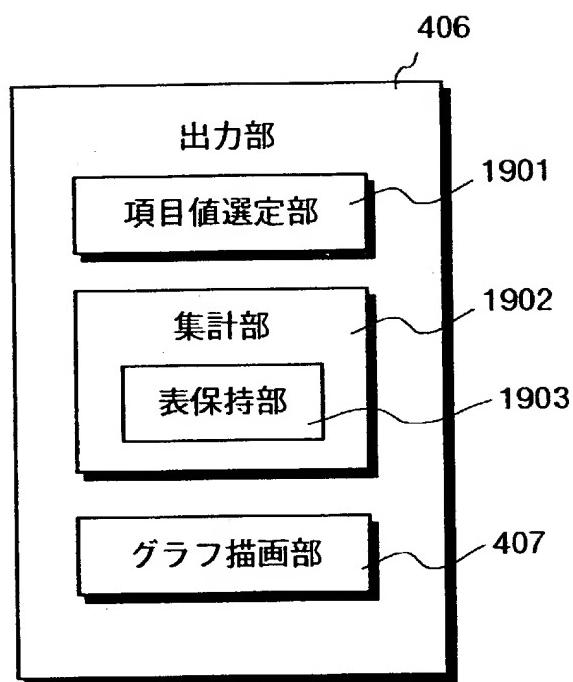
【図17】

1700

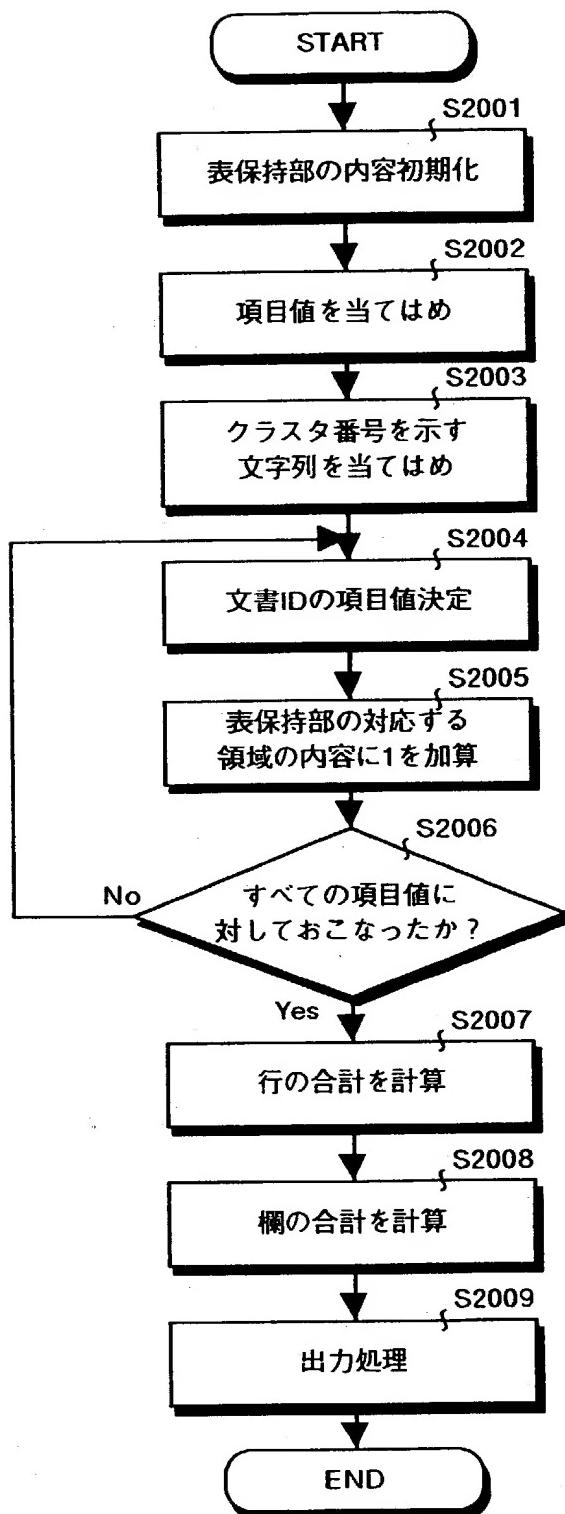
車種	ABC1600	ABC1800	ABC2000	DEF1600	KLM1200	XYZ3000	合計
クラス1	3	0	4	1	1	0	9
クラス2	0	0	2	0	0	23	25
クラス3	5	3	2	7	0	4	21
クラス4	7	6	7	0	0	1	21
.....
合計	227	135	87	194	134	281	1058

【図18】

【図19】



【図20】



【図21】

2101

番号	受付日	車種	年式	内容	△	△
1	1997/3/5	ABC1800	1993	騒音が大きい		
2	1997/3/5	ABC2000	1995	排気声が黒い		
3	1997/3/5	ABC1800	1996	塗装が変色する		
4	1997/3/5	DEF1600	1995	オイル漏れる		
5	1997/3/5	KLM1200	1992	暖房が効かない		
6	1997/3/5	DEF1600	1994	騒音が大きい		
7	1997/3/5	DEF1600	1996	エンジンがかからない		
8	1997/3/5	ABC1800	1997	オーバーヒートが起まる		
9	1997/3/6	XYZ3000	1992	バッテリー上がり		
10	1997/3/6	KLM1200	1993	エンジンがかからない		
11	1997/3/6	ABC1600	1994	排気声が黒い		
12	1997/3/6	DEF1600	1997	ラジオが鳴らない		
13	1997/3/6	ABC1800	1996	塗装がはげる		
14	1997/3/6	XYZ3000	1995	暖房が効かない		

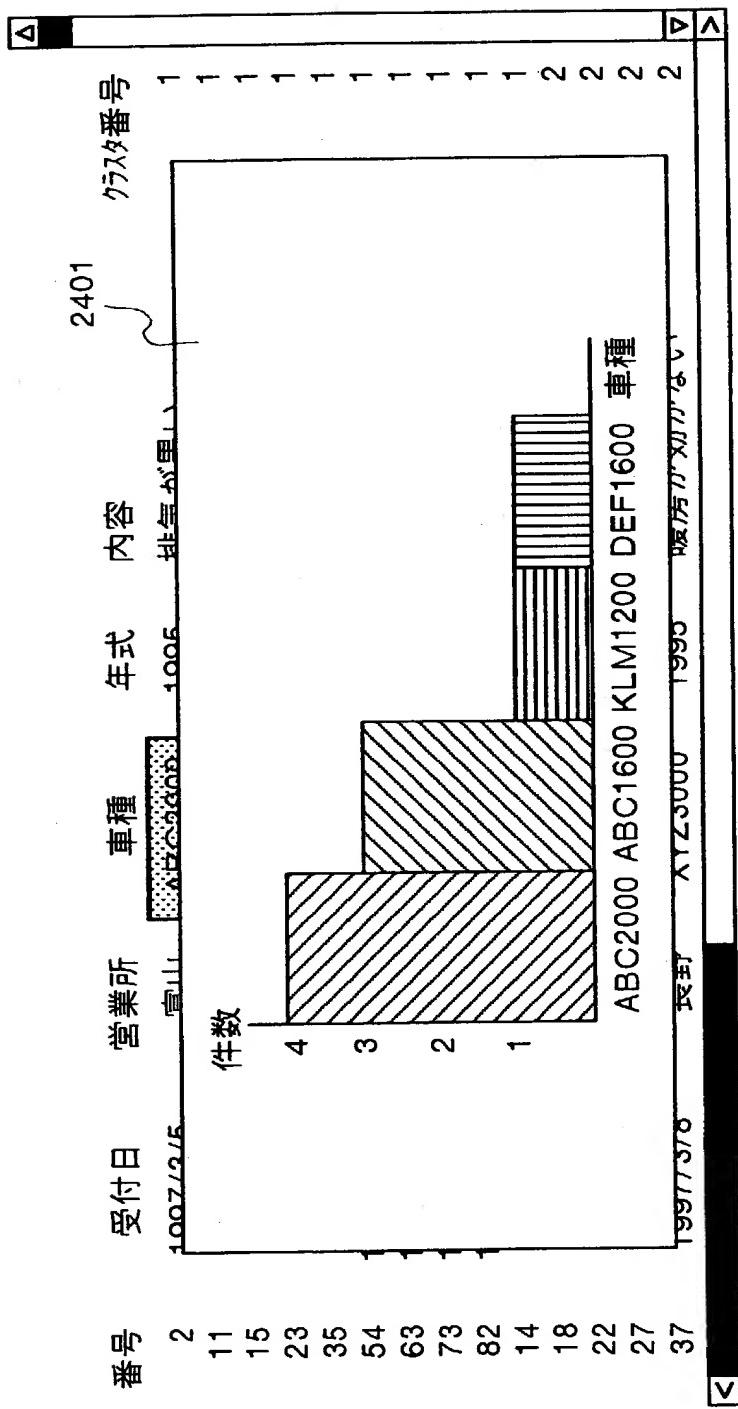
【図22】

番号	受付日	業者所	車種	年式	内容	クラスタ番号
2	1997/3/5	富山	ABC2000	1995	排気が黒い	1
11	1997/3/6	京	ABC1600	1994	排気が黒い	1
15	1997/3/7	幌	ABC2000	1996	排気が黒い	1
23	1997/3/7	岡	ABC2000	1995	排気が黒い	1
35	1997/3/8	野	KLM1200	1992	排気がくさい	1
54	1997/3/10	京	ABC1600	1994	排気が黒い	1
63	1997/3/12	東	ABC2000	1996	排気が黒い	1
73	1997/3/14	長	DEF1600	1997	排気がにおう	1
82	1997/3/14	福	ABC1600	1992	排気が黒い	1
14	1997/3/6	東	XYZ3000	1995	暖房が効かない	1
18	1997/3/7	長	XYZ3000	1997	暖房が効かず、寒い	2
22	1997/3/7	野	XYZ3000	1997	暖房が効かない	2
27	1997/3/8	仙	XYZ3000	1995	暖房の効きが悪い	2
37	1997/3/8	台	XYZ3000	1995	暖房が効かない	2

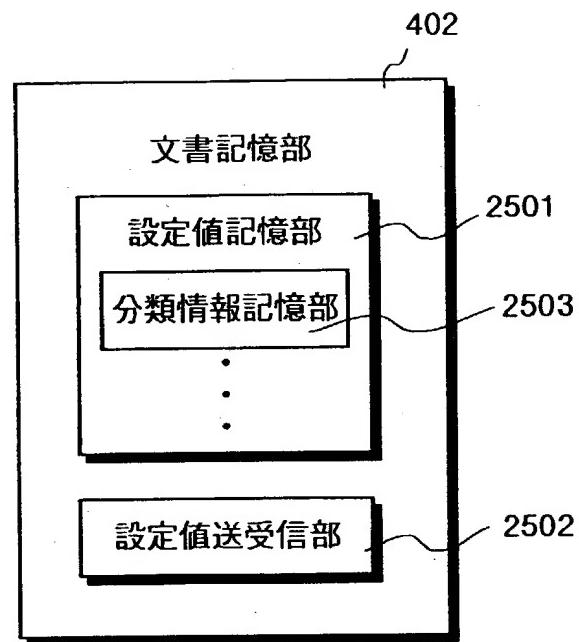
【図23】

番号	受付日	業者所	車種	年式	内容	
2	1997/3/5	富山	ABC2000	1995	排気が黒い	1
11	1997/3/6	京	ABC1600	1994	排気が黒い	1
15	1997/3/7	幌	ABC2000	1996	排気が黒い	1
23	1997/3/7	岡	ABC2000	1995	排気が黒い	1
35	1997/3/8	野	KLM1200	1992	排気がくさい	1
54	1997/3/10	京	ABC1600	1994	排気が黒い	1
63	1997/3/12	京	ABC2000	1996	排気が黒い	1
73	1997/3/14	台	DEF1600	1997	排気がにおう	1
82	1997/3/14	野	ABC1600	1992	排気が黒い	1
14	1997/3/6	仙	XYZ3000	1995	暖房が効かない	1
18	1997/3/7	長	XYZ3000	1997	暖房が効かず、寒い	2
22	1997/3/7	福	XYZ3000	1997	暖房が効かない	2
27	1997/3/8	東	XYZ3000	1995	暖房の効きが悪い	2
37	1997/3/8	長	XYZ3000	1995	暖房が効かない	2

【図24】



【図25】



【図26】

ファイル番号	登付日	営業所	車種	年式	内容
2605	1997/3/6	ABC1800	1996	排気ガス漏れする	
2603	1997/3/10	ABC1800	1994	排気ガス漏れする	
2602	1997/3/12	ABC1800	1993	排気ガス漏れする	
2604	1997/3/17	ABC1800	1993	排気ガス漏れする	
2601	1997/3/20	ABC1800	1996	排気ガス漏れする	
2601	1997/3/22	ABC1800	1995	排気ガス漏れする	
2601	1997/3/24	ABC1800	1996	排気ガス漏れする	
2601	1997/3/24	ABC1800	1994	排気ガス漏れする	
2601	1997/3/5	DEF1600	1995	排気ガス漏れする	
53	分類数を入力してください	50			
1	分類数 : 50				
8					
13	1997/3/6	盛岡	ABC1800	1996	排気ガス漏れする
18	1997/3/10	大阪	ABC1800	1994	排気ガス漏れする
28	1997/3/12	広島	ABC1800	1995	排気ガス漏れする
35	1997/3/17	横浜	ABC1800	1996	排気ガス漏れする
39	1997/3/20	愛知	ABC1800	1993	排気ガス漏れする
42	1997/3/22	東京	ABC1800	1996	排気ガス漏れする
46	1997/3/24	富山	ABC1800	1997	排気ガス漏れする
47	1997/3/24	大阪	ABC1800	1994	排気ガス漏れする
4	1997/3/5	札幌	DEF1600	1995	排気ガス漏れする

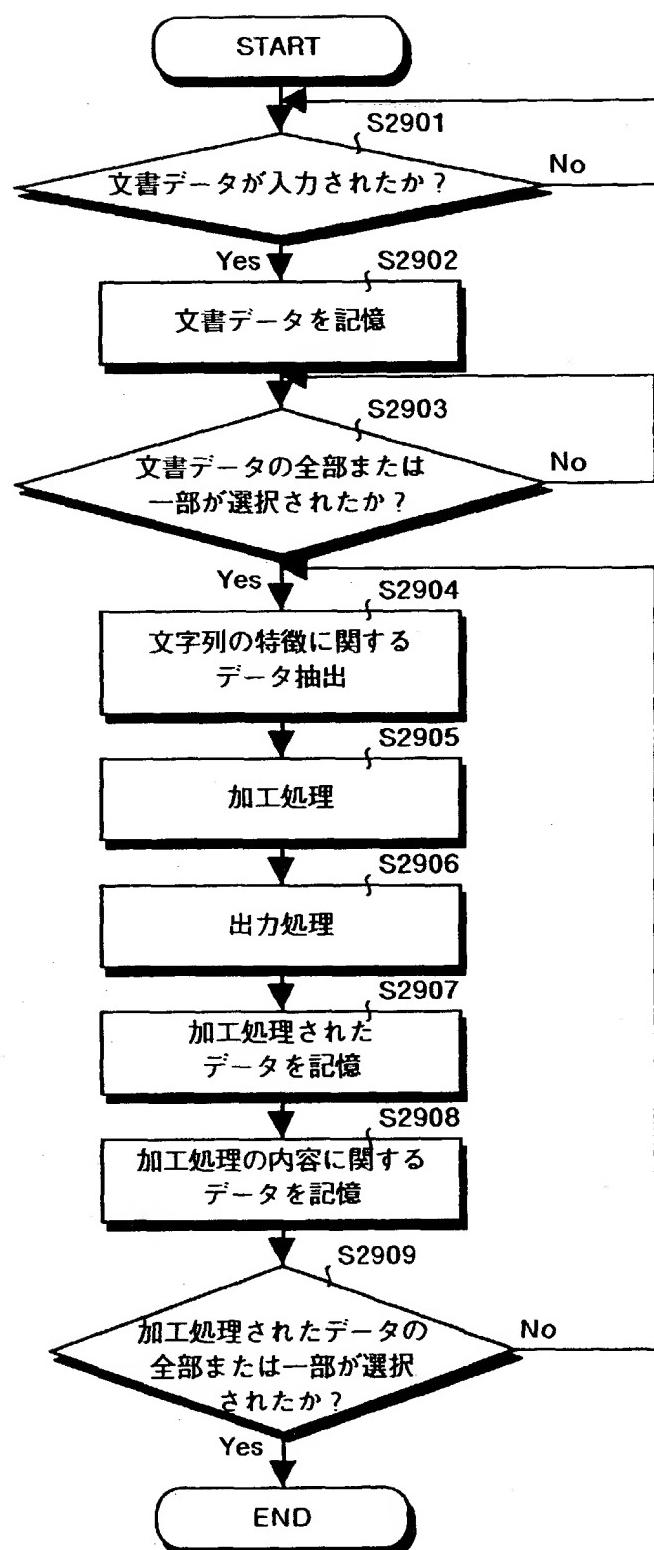
【図27】

番号	受付日	業者所	車種	年式	内容	クラスター番号 ▲	
						2701	2702
1	1997/3/5	愛知	ABC18	分類日時 :	1997年5月12日	5	1
2	1997/3/5	富山	ABC20		16時32分	1	7
3	1997/3/5	東京	ABC18	分類対象数 :	823	11	11
4	1997/3/5	幌岡	DEF18			2	2
5	1997/3/5	別	KLM12			5	5
6	1997/3/5	福	DEF18	<分類設定値>		8	8
7	1997/3/5	長野	DEF18	分類数 :	30	1	1
8	1997/3/5	高松	ABC18	分類品詞 :	名詞	14	14
9	1997/3/6	東京	XYZ30			12	12
10	1997/3/6	崎	KLM1200		1993 エンブレムがつかない	8	8
11	1997/3/6	大阪	ABC1600		1994 排氣が黒い	1	1
12	1997/3/6	長野	DEF1600		1997 ラジオが鳴らない	6	6
13	1997/3/6	岡	ABC1800		1996 塗装がはげる	7	7
14	1997/3/6	仙台	XYZ3000		1995 暖房が効かない	2	2

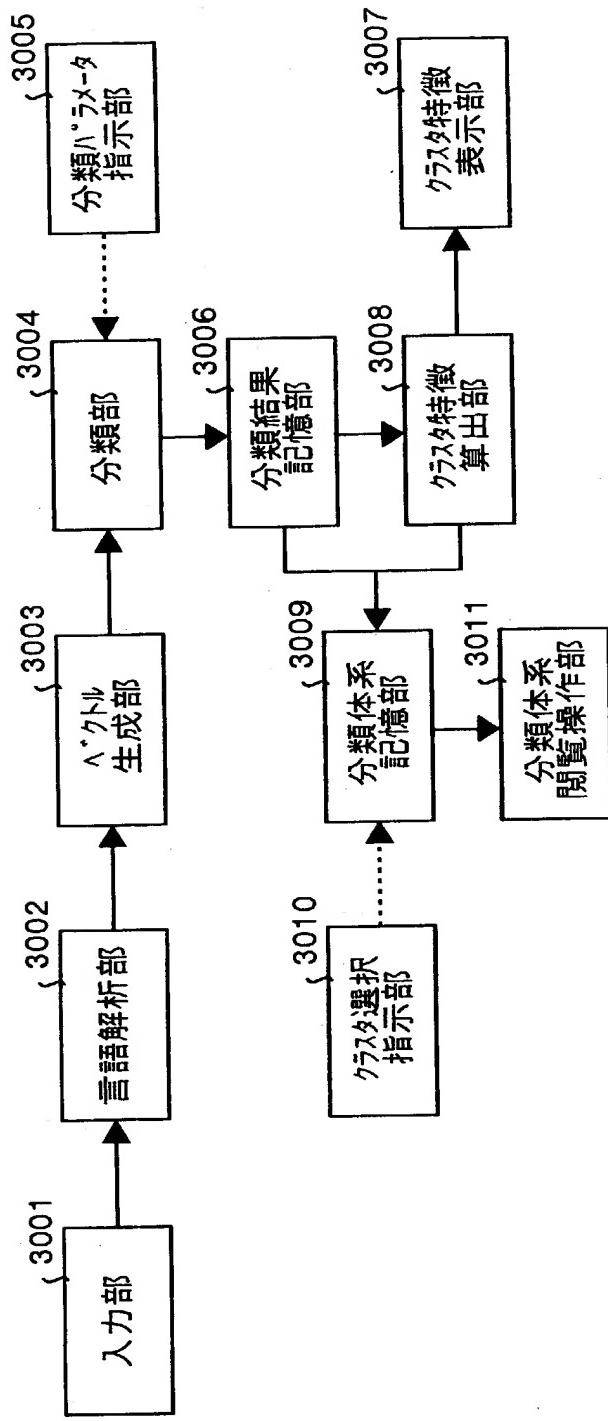
【図28】

番号	受付日	営業所	車種	年式	内容	クラス番号
1	1997/3/5	愛知	ABC1800	1993	騒音が大きい	5
2	1997/3/5	富山	ABC2000	1995	排気が黒い	1
3	1997/3/5	東京	ABC1800	1996	塗装が変色する	7
4	1997/3/5	幌岡	DEF1600	1995	オイルが漏れる	11
5	1997/3/5	別	KLM1200	1992	暖房が効かない	2
6	1997/3/5	札幌	DEF1600	1994	騒音が大きい	5
7	1997/3/5	長野	DEF1600	1996	エンジンがかからない	8
8	1997/3/5	東京	ABC1800	1997	オーバーヒートが起こる	14
9	1997/3/6	松	XYZ3000	1992	バッテリーが上がる	12
10	1997/3/6	崎	KLM1200	1993	エンジンがかからない	8
11	1997/3/6	大阪	ABC1600	1994	排気が黒い	1
12	1997/3/6	長	DEF1600	1997	ラジオが鳴らない	6
13	1997/3/6	阪	ABC1800	1996	塗装がはげる	7
14	1997/3/6	盛岡	XYZ3000	1995	暖房が効かない	2
					2801	
					<分類設定値>分類数：30 分類品詞：名詞	
					分類結果1	分類結果2

【図29】



【図30】

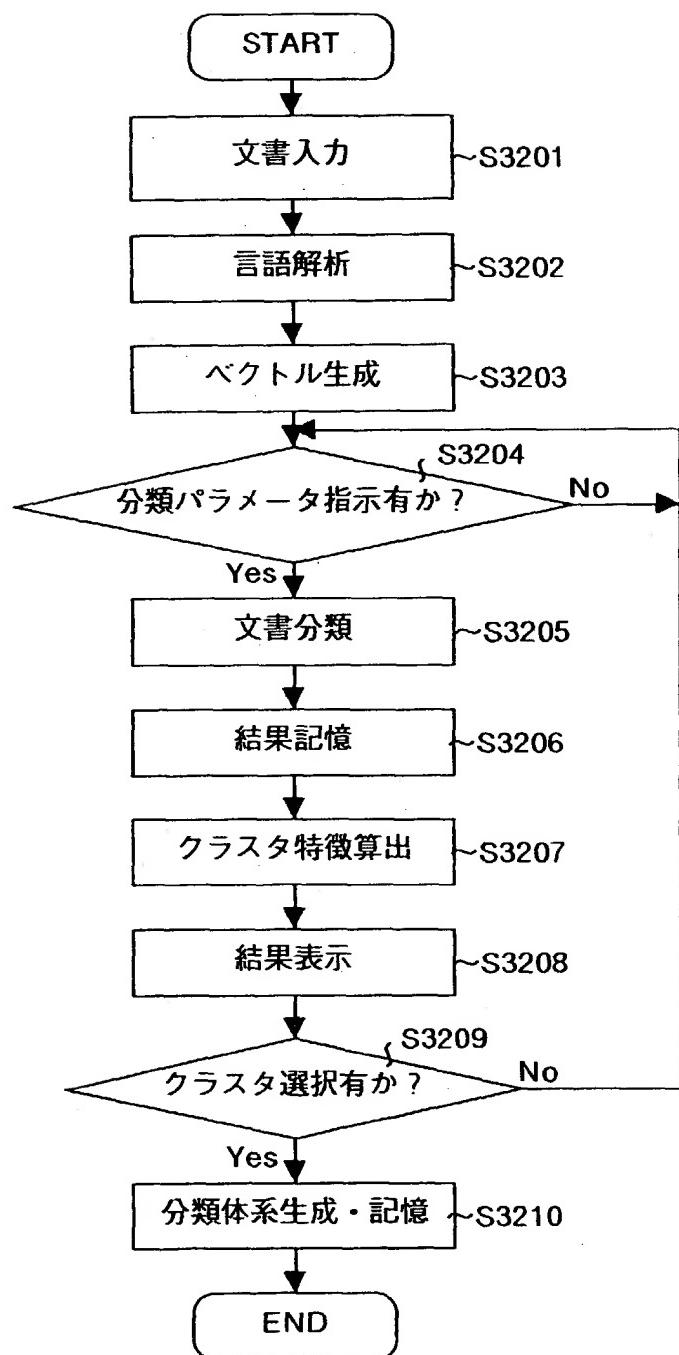


【図31】

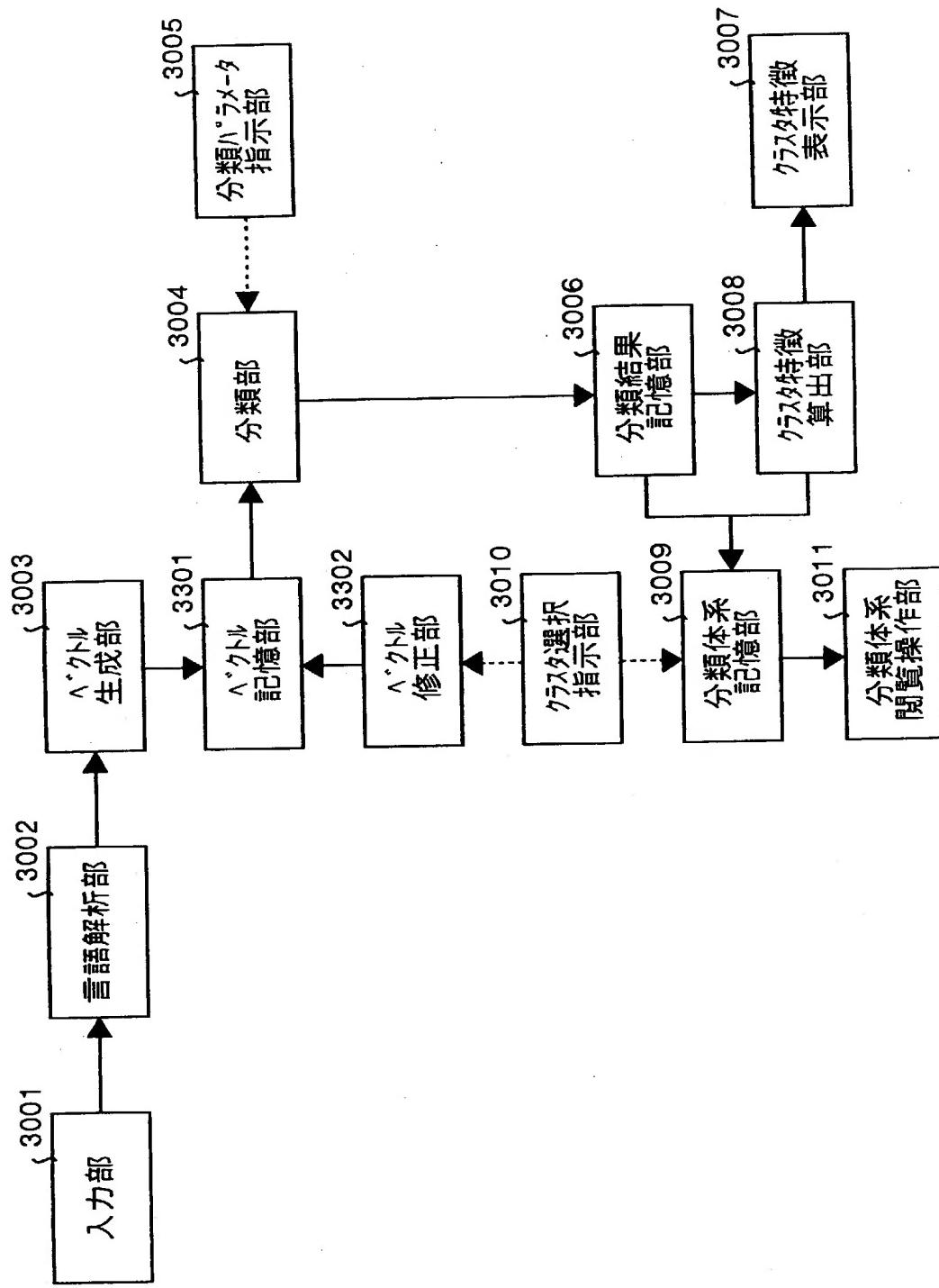
クラスID	メンバーカー数	頻度の高い単語	文書内容	重心との類似度
1	248	管理者、多忙…	システム管理業務が多くて多忙だ	0.987
			システム管理が多忙でコスト削減できない	0.965
			管理者が多忙だとシステムがタ'ンする	0.911
			システムタ'ンで管理が大変	0.889
			システムタ'ンで管理業務部が多忙になる	0.876

N	1498	操作性、悪い…	ソフトの操作性が悪い、	0.969
			ソフトの操作性を覚えるの大変である	0.962

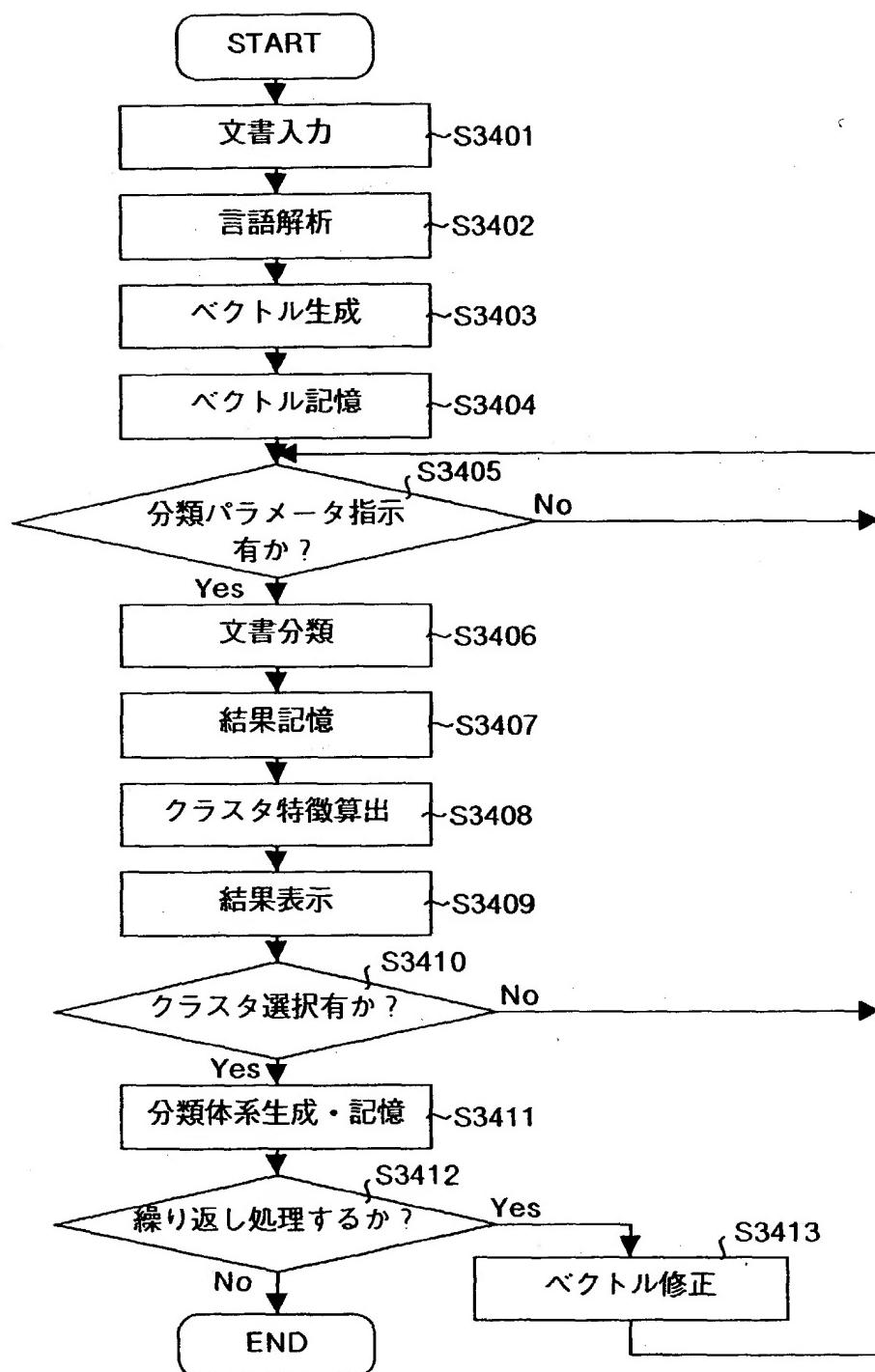
【図32】



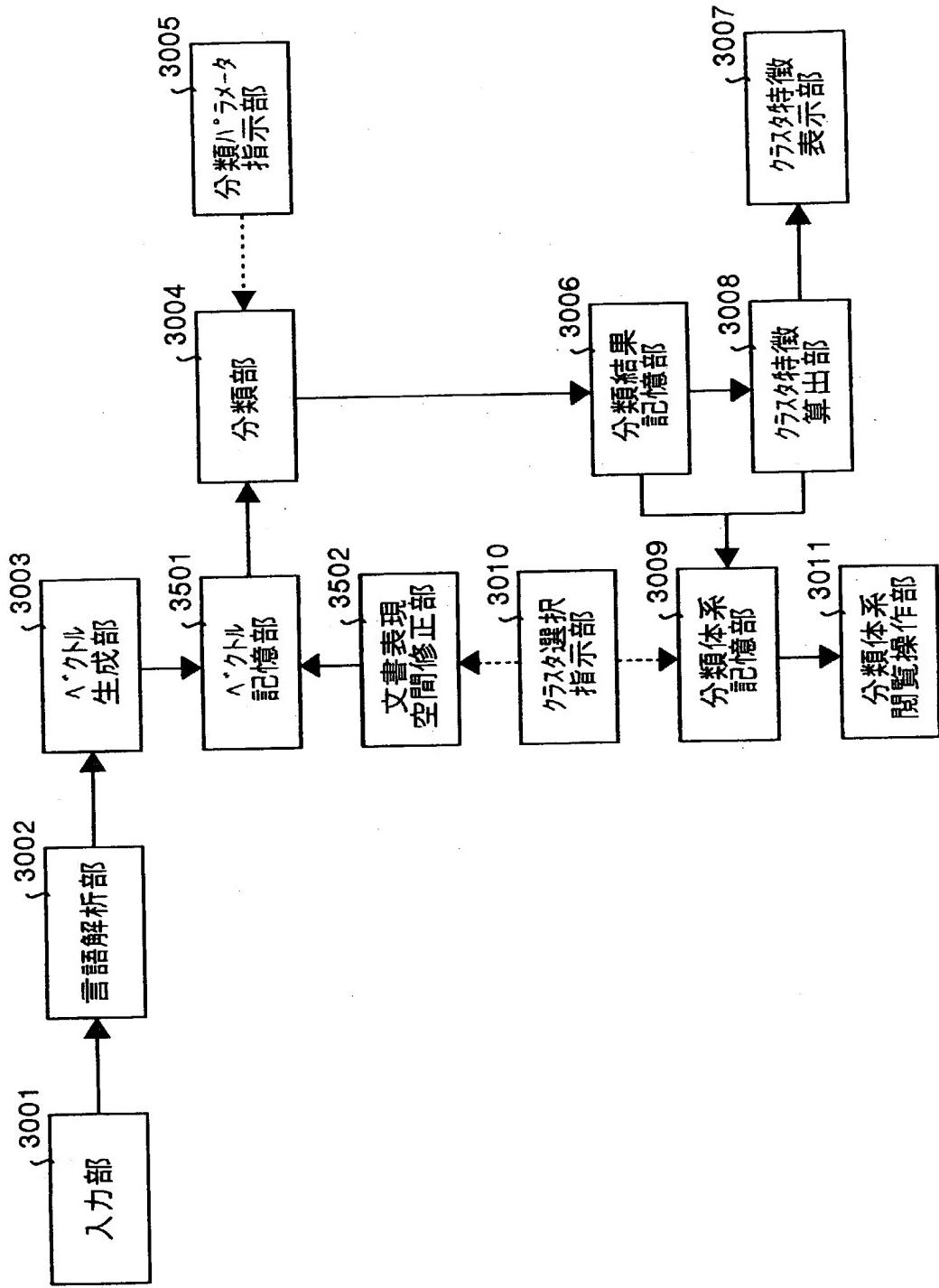
【図33】



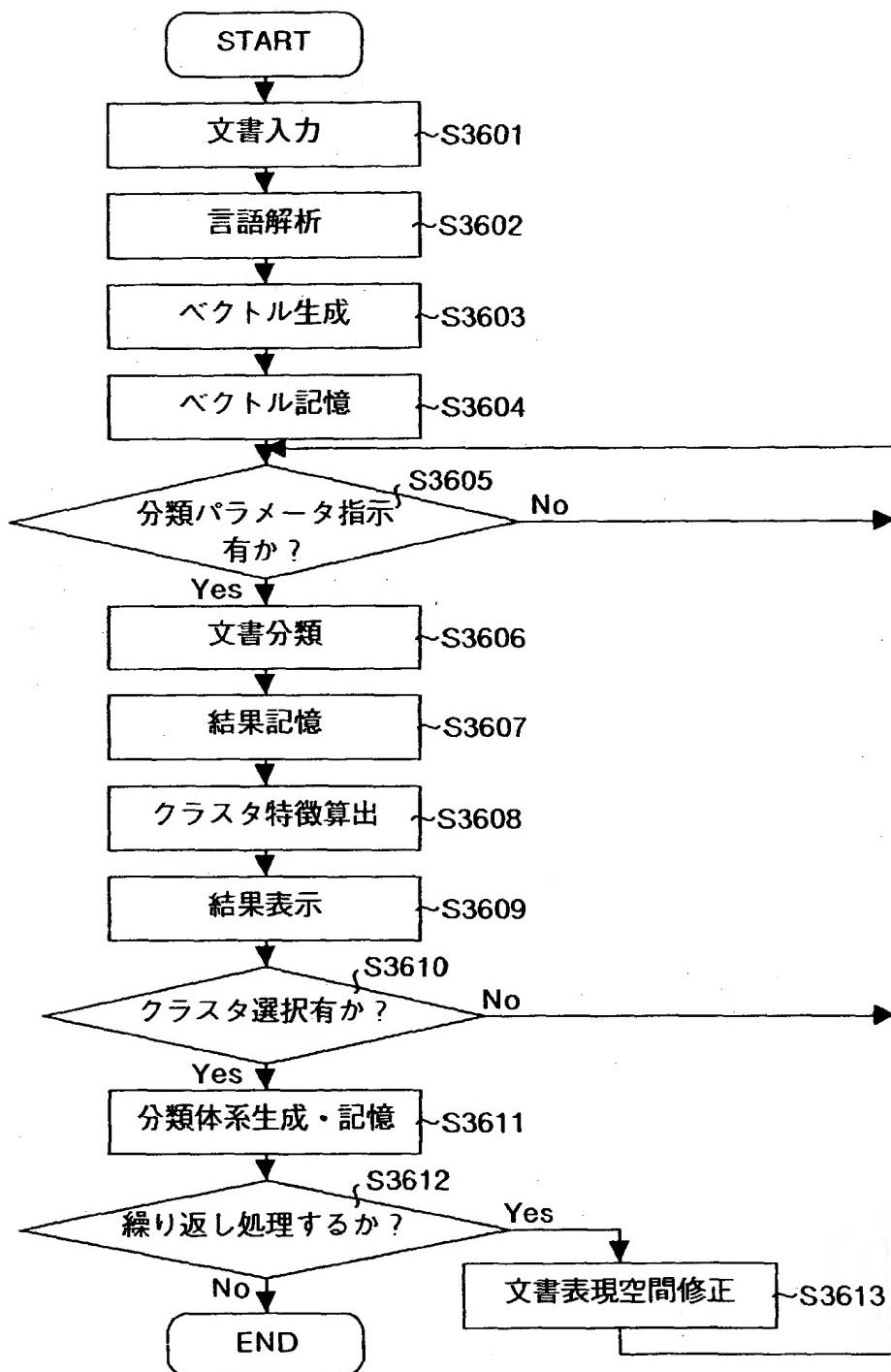
【図34】



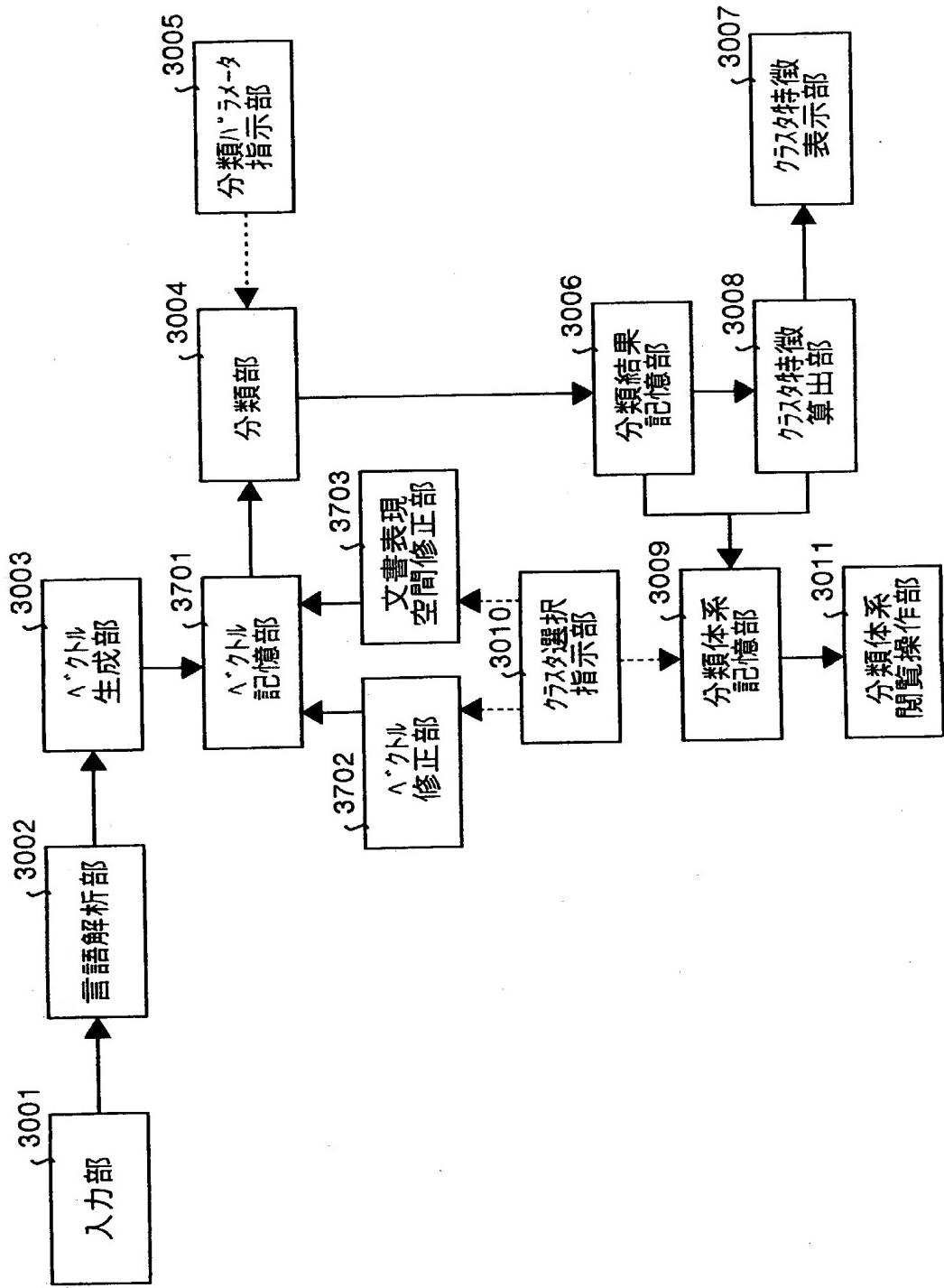
【図35】



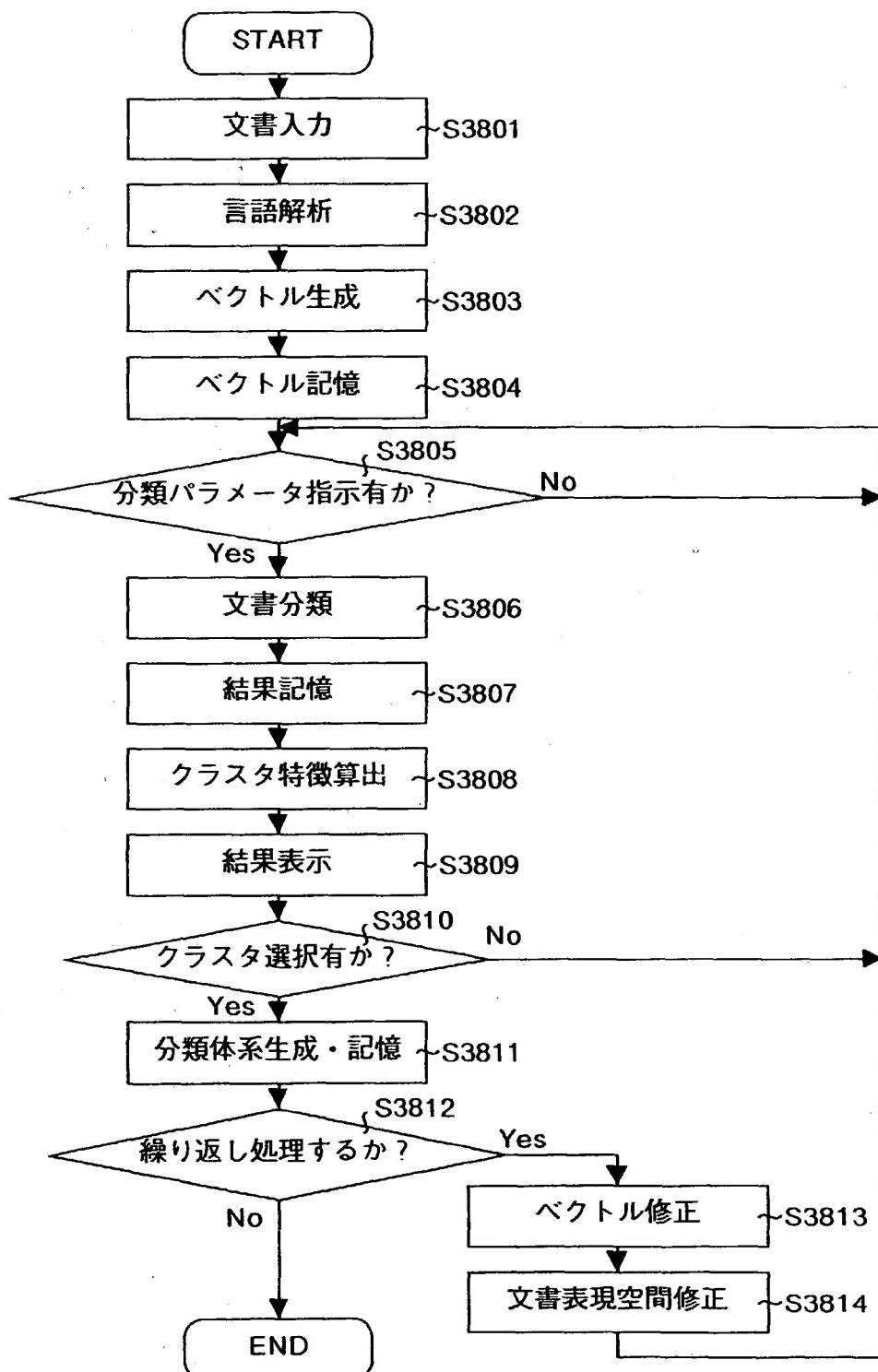
【図36】



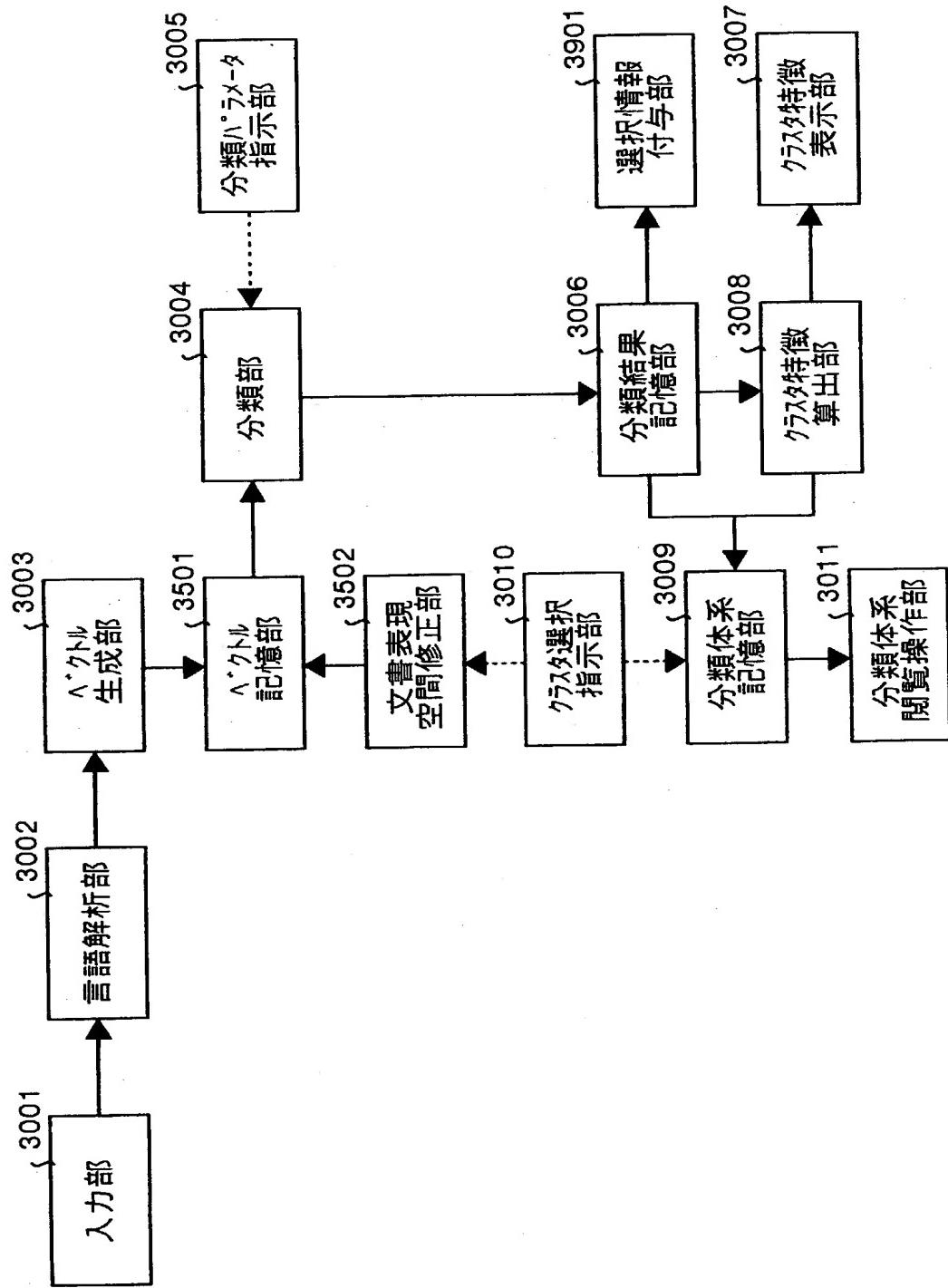
【図37】



【図38】



【図39】

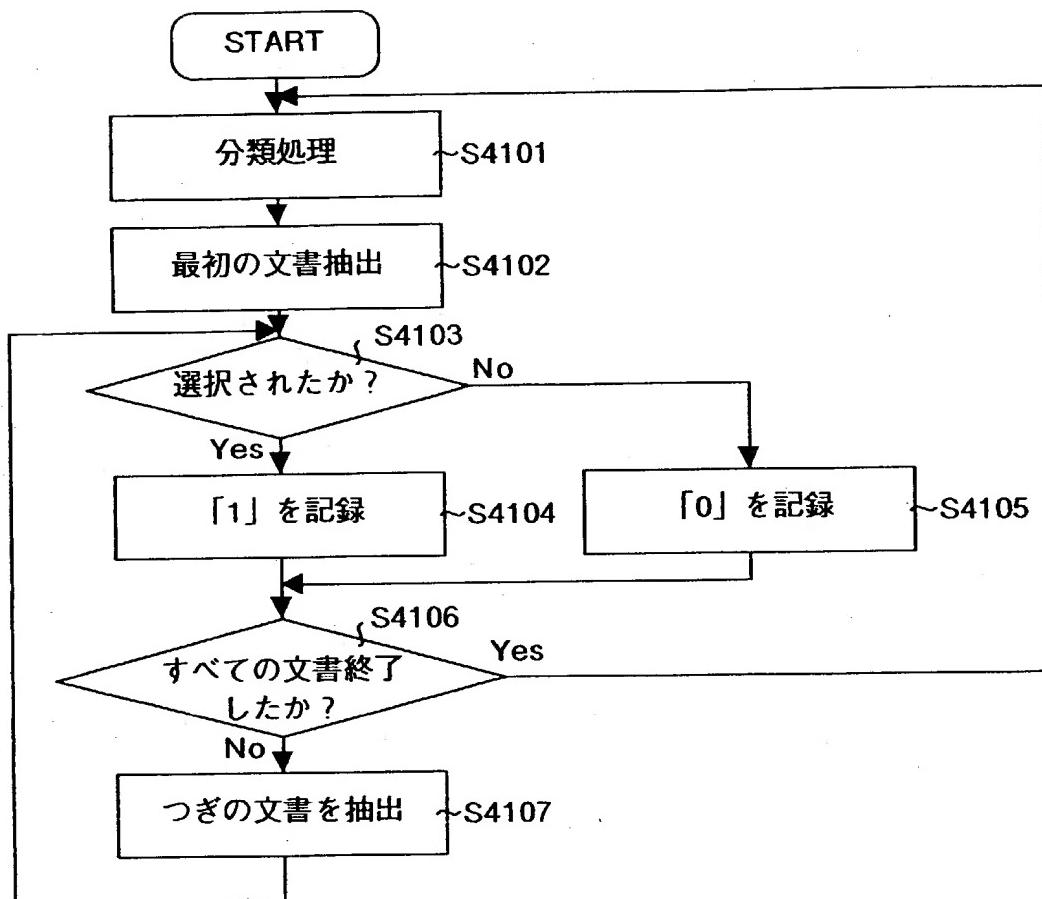


【図40】

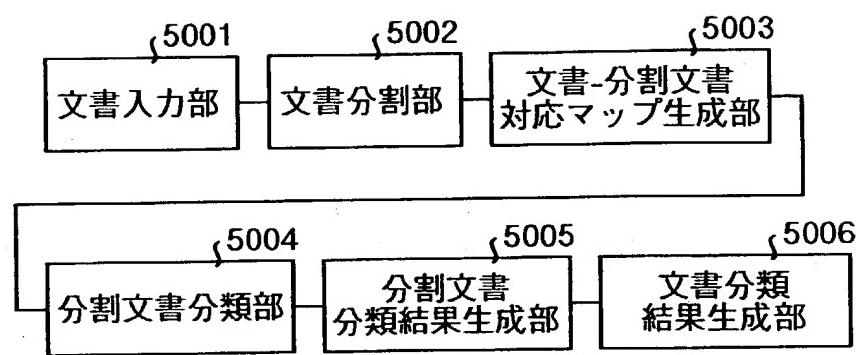
↙
4000

選択情報 (1 = 選択 / 0 = 未選択) 分類回数順	
文書ID	
1	[1, 1, 0, 0]
2	[0, 0, 0, 0]
3	[0, 1, 0, 0]
⋮	⋮
n-1	[1, 0, 0, 0]
n	[0, 0, 0, 0]

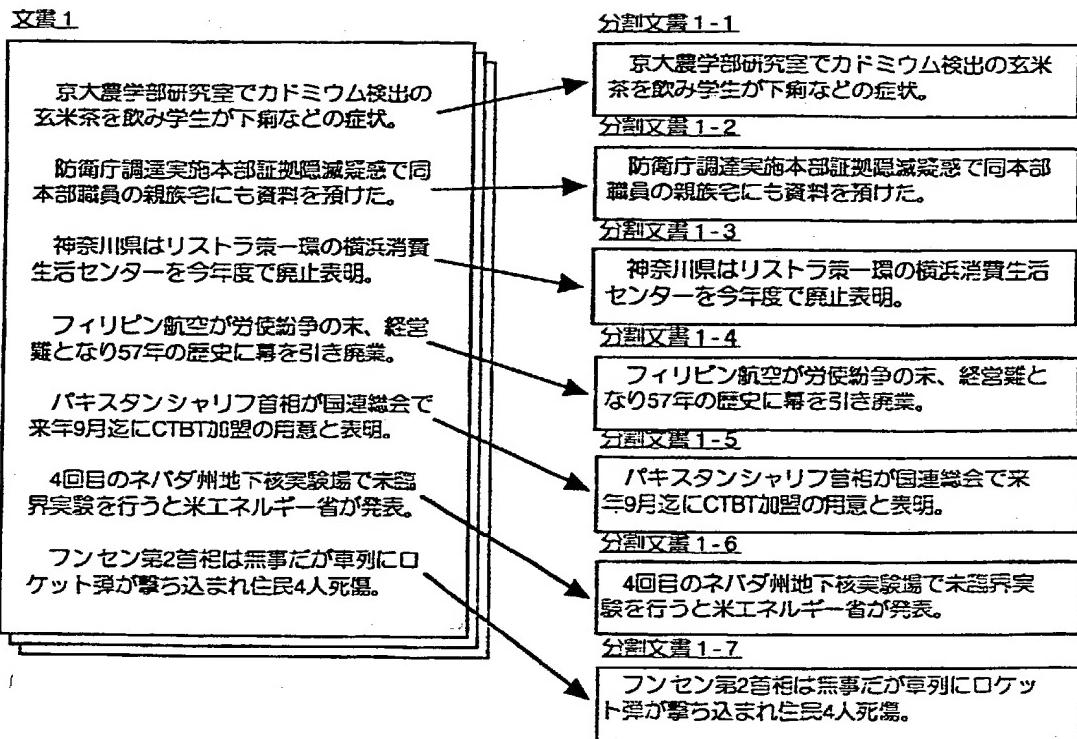
【図4-1】



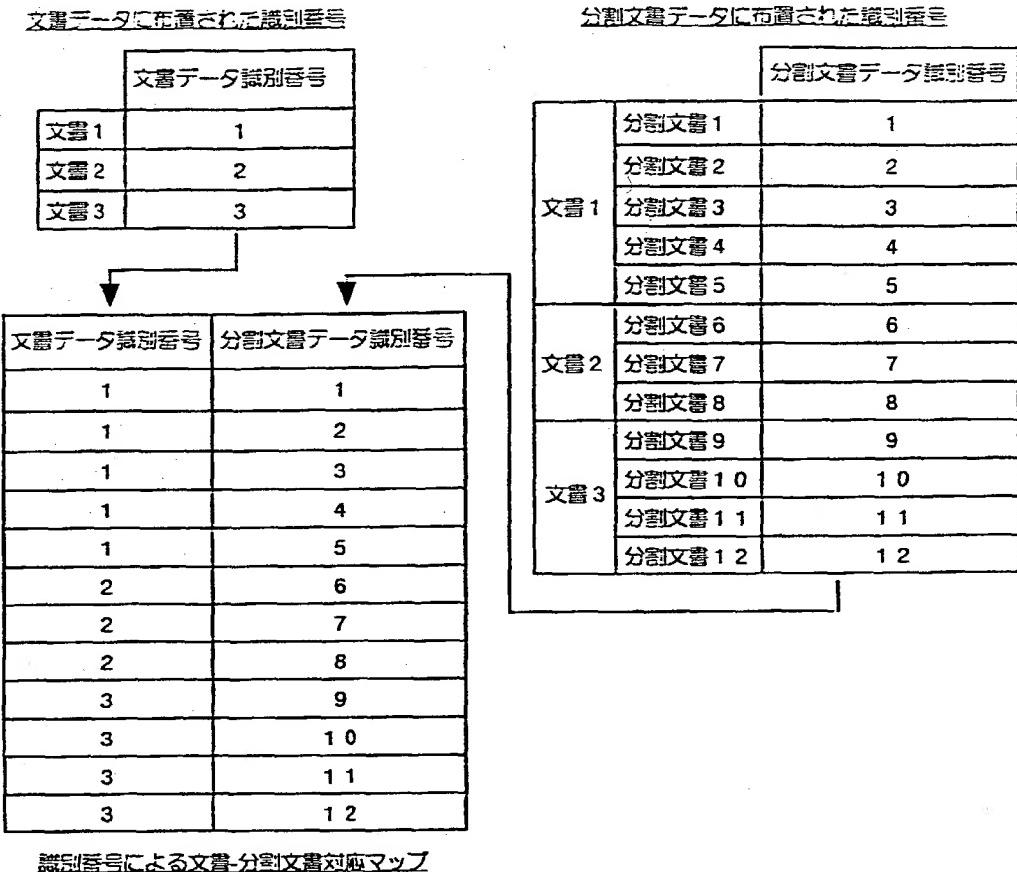
【図42】



【図43】



【図44】



【図45】

分割文書データの特徴ベクトル表現

分割文書データ識別番号	分割文書データ特徴ベクトル
分割文書1	(1, 1, 1)
分割文書2	(5, 5, 5)
分割文書3	(3, 2, 4)
分割文書4	(3, 2, 3)
分割文書5	(5, 4, 6)
分割文書6	(1, 2, 1)
分割文書7	(1, 0, 1)
分割文書8	(5, 4, 5)
分割文書9	(2, 2, 4)
分割文書10	(2, 1, 1)
分割文書11	(4, 4, 6)
分割文書12	(5, 5, 6)

文書分類（クラスタ分析手法を適用）

分割文書データを3つのカテゴリに分類した結果

分割文書データ識別番号	分類カテゴリ	所属カテゴリの代表値との距離
分割文書1	カテゴリ1	0.25
分割文書2	カテゴリ3	0.87
分割文書3	カテゴリ2	0.48
分割文書4	カテゴリ2	0.74
分割文書5	カテゴリ3	0.54
分割文書6	カテゴリ1	1.03
分割文書7	カテゴリ1	1.03
分割文書8	カテゴリ3	0.70
分割文書9	カテゴリ2	0.74
分割文書10	カテゴリ1	0.75
分割文書11	カテゴリ3	0.94
分割文書12	カテゴリ3	0.83

分類カテゴリに関する情報

	代表値（所属分割文書データの重心）	所属データ数
カテゴリ1	(1.25, 1.00, 1.00)	4
カテゴリ2	(2.66, 2.00, 3.66)	3
カテゴリ3	(4.80, 4.40, 5.60)	5

分類カテゴリ間の距離

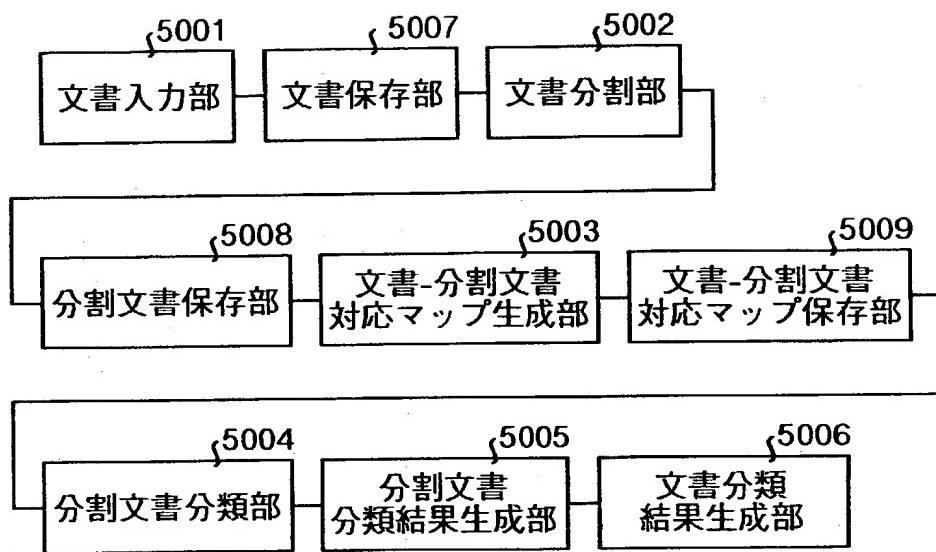
	カテゴリ2	カテゴリ3
カテゴリ1	3.17	6.68
カテゴリ2		3.69

【図46】

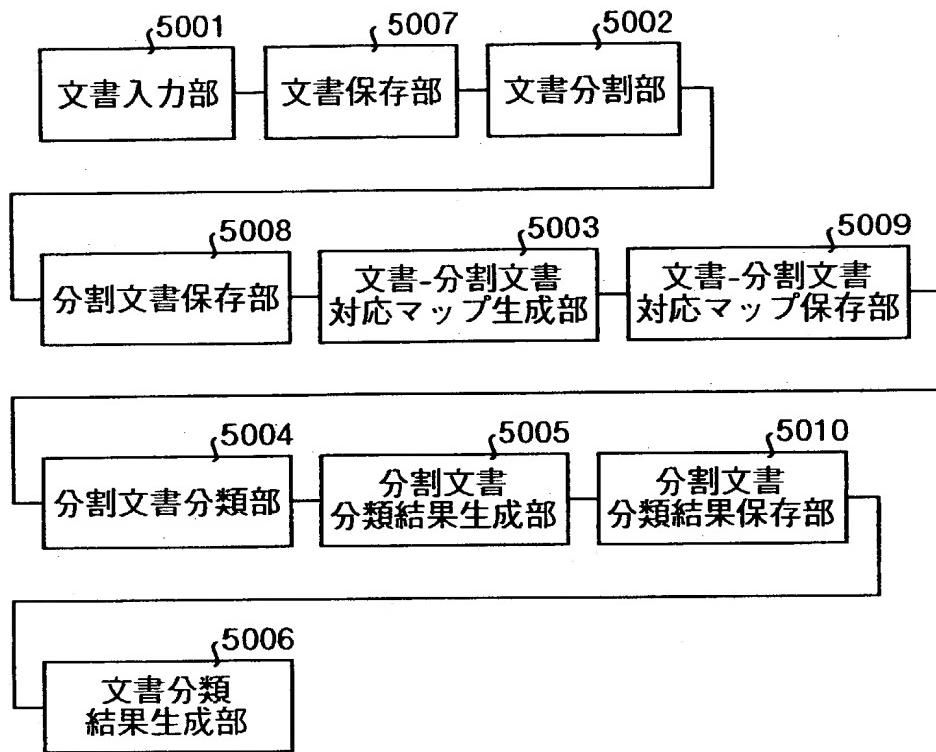
文書分類結果

分類カテゴリ	分割文書	類似度	所属文書	文書占有率	相対位置	類似順位
カテゴリ1	分割文書1	0.25	文書1	1/5	1/5	1
カテゴリ1	分割文書6	1.03	文書2	2/3	1/3	3
カテゴリ1	分割文書7	1.03	文書2	2/3	2/3	3
カテゴリ1	分割文書10	0.75	文書3	1/4	2/4	2
カテゴリ2	分割文書3	0.48	文書1	2/5	3/5	1
カテゴリ2	分割文書4	0.74	文書1	2/5	4/5	2
カテゴリ2	分割文書9	0.74	文書3	1/4	1/4	2
カテゴリ3	分割文書2	0.87	文書1	2/5	2/5	4
カテゴリ3	分割文書5	0.54	文書1	2/5	5/5	1
カテゴリ3	分割文書8	0.70	文書2	1/3	3/3	2
カテゴリ3	分割文書11	0.94	文書3	2/4	3/4	4
カテゴリ3	分割文書12	0.83	文書3	2/4	4/4	3

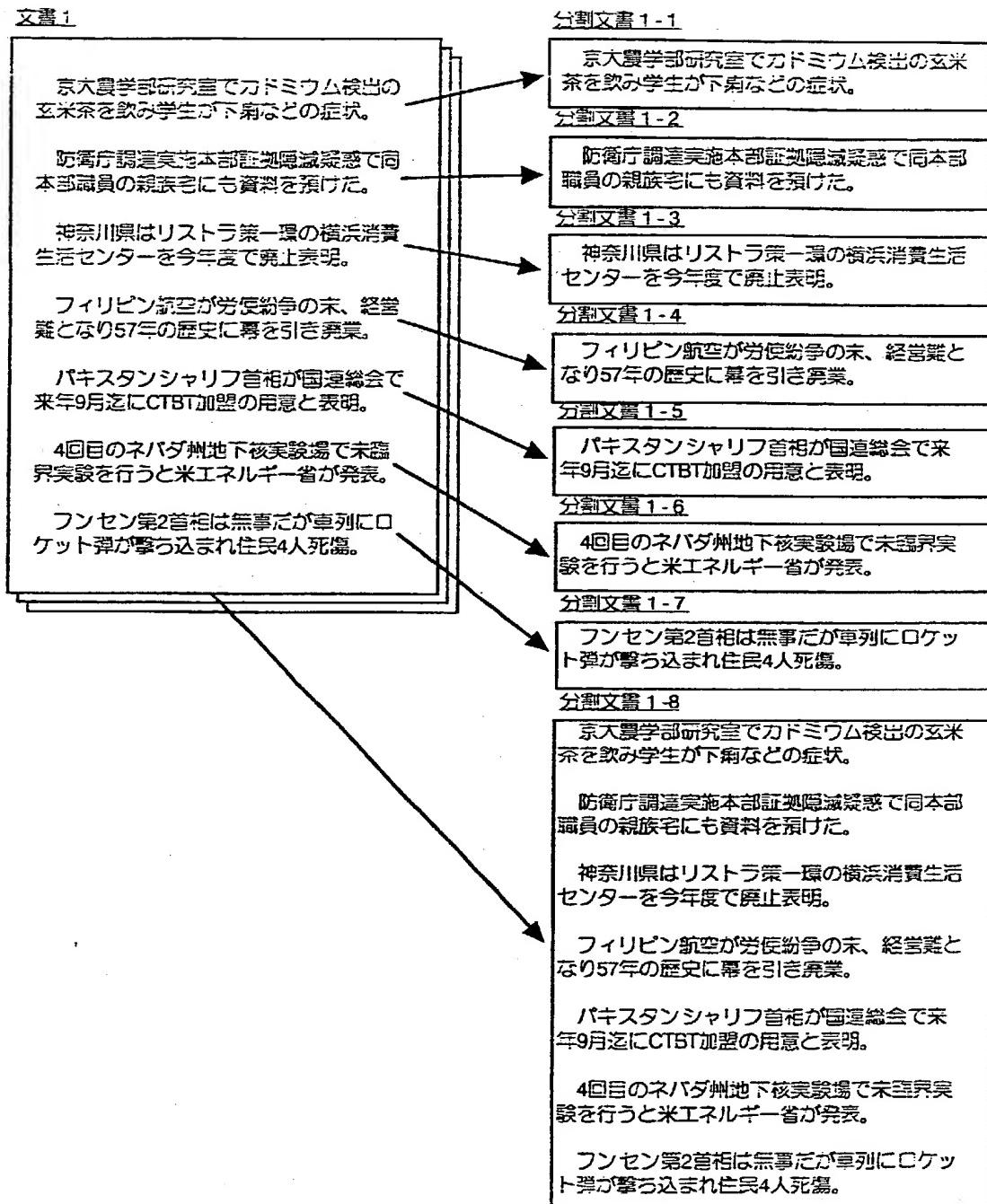
【図47】



【図48】



【図49】



【図50】

文書データ

ニューストピック (98/09/25)

- ・京大農学部研究室でカドミウム検出の玄米茶を飲み学生が下痢などの症状。
- ・防衛庁調査実施本部証拠隠滅疑惑で同本部職員の親族宅にも資料を預けた。
- ・神奈川県はリストラ策一環の横浜消費生活センターを今年度で廃止表明。
- ・フィリピン航空が労使紛争の末、経営難となり57年の歴史に幕を引き廃業。
- ・パキスタンシャリフ首相が国連総会で来年9月迄にCTBT加盟の用意と表明。
- ・4回目のネバダ州地下核実験場で未臨界実験を行うと米エネルギー省が発表。
- ・フンセン第2首相は無事だが車列にロケット弾が撃ち込まれ住民4人死傷。

HTML形式

```

<HTML>
<HEAD>
<META NAME=GENERATOR CONTENT="Claris Home Page 2.0">
<META HTTP-EQUIV="Content-Type" CONTENT="text/html;CHARSET=x-sjis">
<X-SAS-WINDOW TCP=543 CTOM=769 LEFT=224 RIGHT=656>
</HEAD>
<BODY>

<P> </P>

<P> ニューストピック (98/09/25) </P>

<UL>
<LI>京大農学部研究室でカドミウム検出の玄米茶を飲み学生が下痢などの症状。
<LI>防衛庁調査実施本部証拠隠滅疑惑で同本部職員の親族宅にも資料を預けた。
<LI>神奈川県はリストラ策一環の横浜消費生活センターを今年度で廃止表明。
<LI>フィリピン航空が労使紛争の末、経営難となり57年の歴史に幕を引き廃業。
<LI>パキスタンシャリフ首相が国連総会で来年9月迄にCTBT加盟の用意と表明。
<LI>4回目のネバダ州地下核実験場で未臨界実験を行うと米エネルギー省が発表。
<LI>フンセン第2首相は無事だが車列にロケット弾が撃ち込まれ住民4人死傷。
</UL>
</BODY>
</HTML>

```

文書データの分割

|| タグを持つテキストをひとつの分割文書データとする ||

分割文書データ

分割文書 1

- ・京大農学部研究室でカドミウム検出の玄米茶を飲み学生が下痢などの症状。

分割文書 2

- ・防衛庁調査実施本部証拠隠滅疑惑で同本部職員の親族宅にも資料を預けた。

分割文書 3

- ・神奈川県はリストラ策一環の横浜消費生活センターを今年度で廃止表明。

分割文書 4

- ・フィリピン航空が労使紛争の末、経営難となり57年の歴史に幕を引き廃業。

分割文書 5

- ・パキスタンシャリフ首相が国連総会で来年9月迄にCTBT加盟の用意と表明。

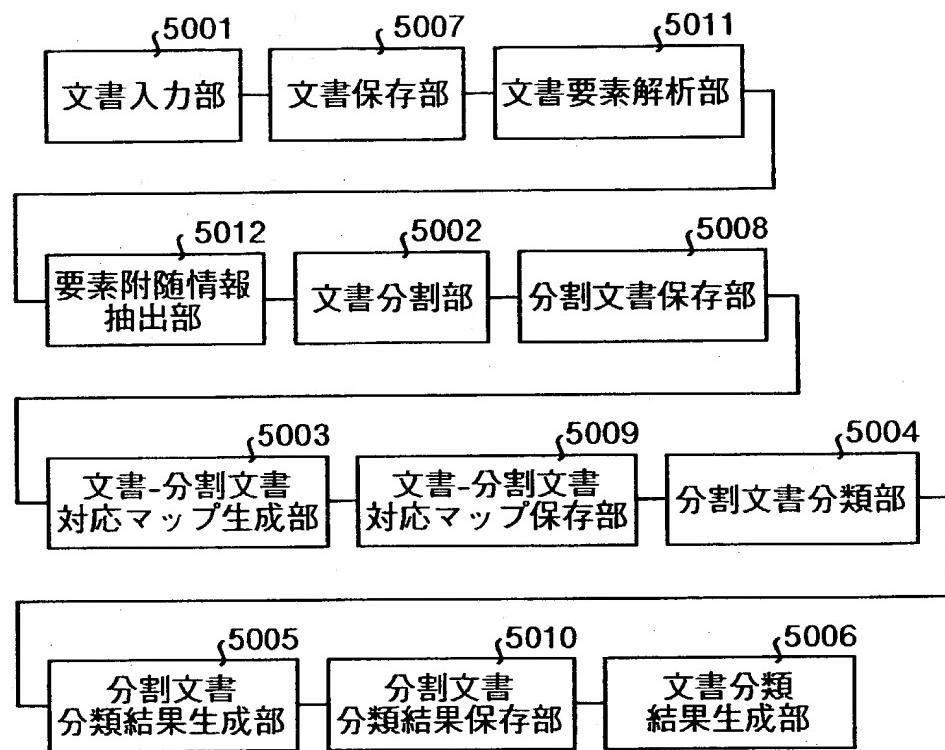
分割文書 6

- ・4回目のネバダ州地下核実験場で未臨界実験を行うと米エネルギー省が発表。

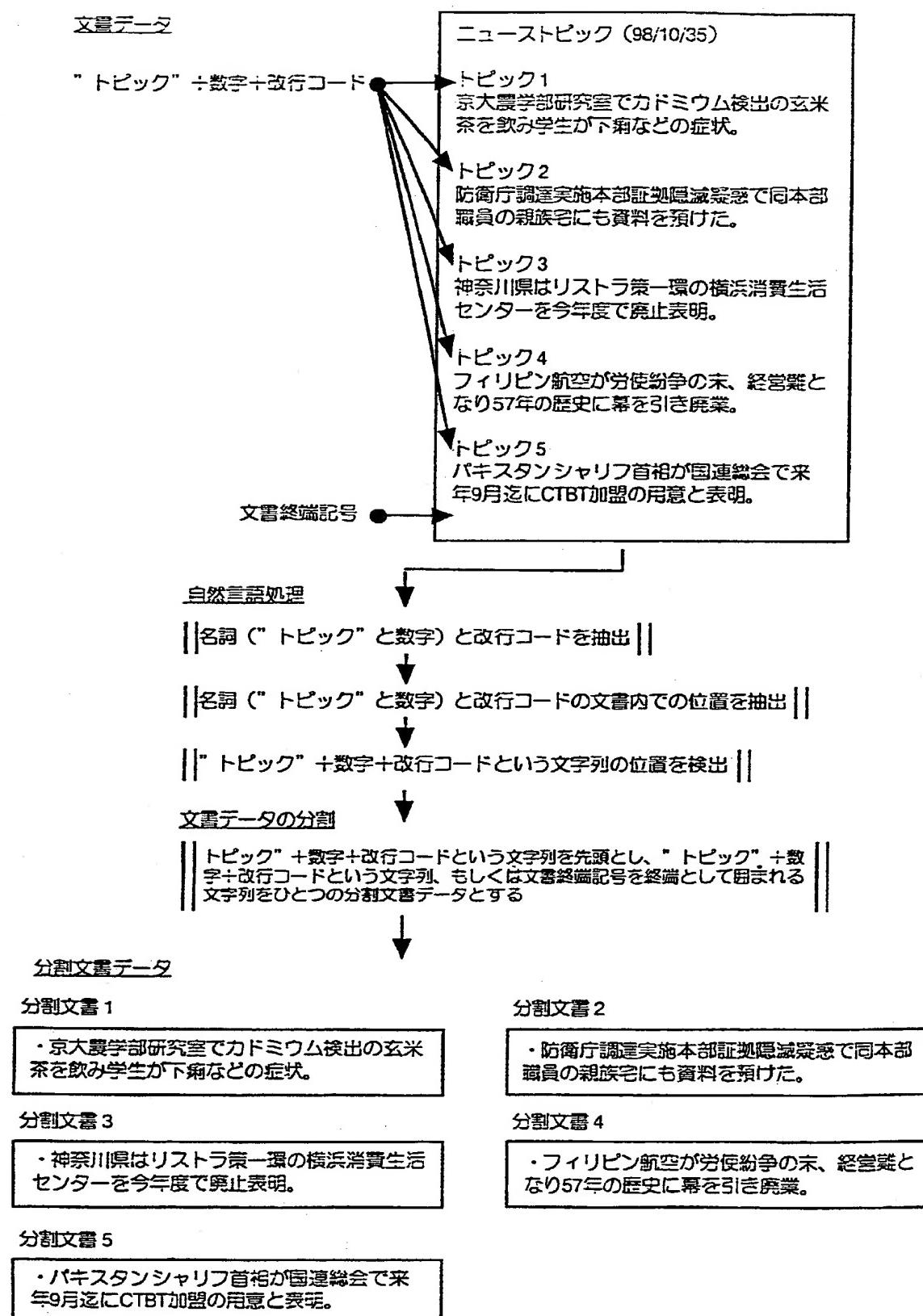
分割文書 7

- ・フンセン第2首相は無事だが車列にロケット弾が撃ち込まれ住民4人死傷。

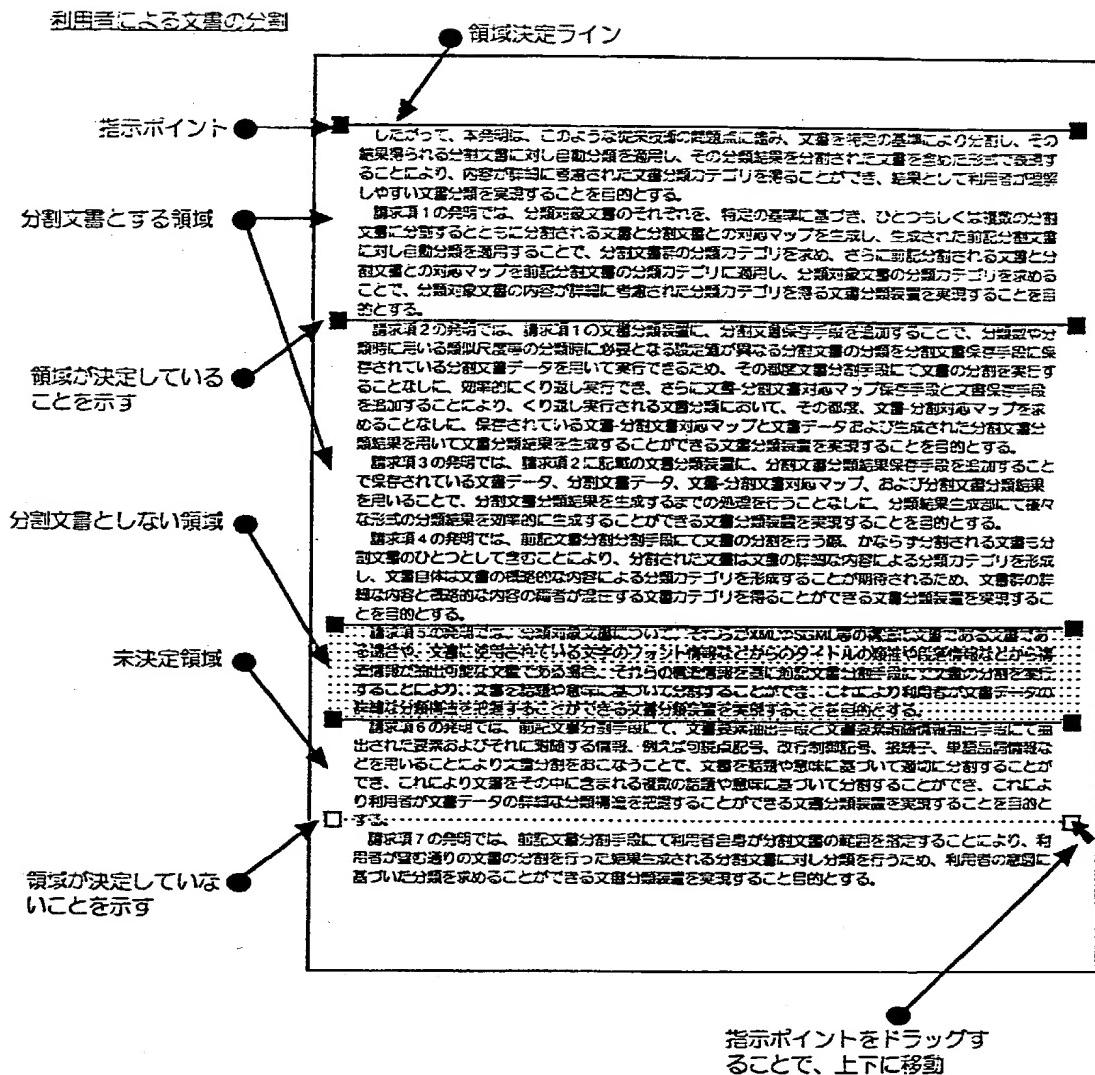
【図 5 1】



【図52】



【図53】



【図54】

文字数と文数による分割

文書データ

ユーザーの要因を反映するような文書分類をおこなうためのひとつの方針として、前記表現空間変換関数により構成される空間における不必要的特徴次元や、悪影響を及ぼすような特徴次元に対し削除や合成をおこなったり、逆にある特徴次元を強調させるための操作をすることが考えられる。しかし、前記表現空間変換関数により生成される空間の特徴次元は、前記文書解析部にて抽出される単語のうち意味的に似たものが複数結合したものと考えることができるため、ユーザーに各特徴次元の意味を提示することは極めて難しい。そこで、ユーザーに分類に反映させたくない内容や強調したい内容をもつ文書や單語などの情報を指定させ、それらを前記表現空間変換関数により構成される空間に適切に射影し、それらと類似度の高い特徴次元や低い特徴次元を判別することで、操作をおこなう特徴次元を選択することが考えられる。ここでは、前記表現空間変換関数の特徴次元を操作する例として、ユーザーが指定するある文書と類似度の高い特徴次元の削除を行う例を示す。ユーザーにより指定された文書を前記文書特徴ベクトルと同じ次元数をもつベクトルで表現し、その文書ベクトルに前記表現空間変換関数を適用し文書ベクトルを前記表現空間変換関数により構成される空間へ射影する。そして、この射影された文書ベクトルと各特徴次元との類似度を算出することで、類似度の高い特徴次元を判別する。このとき、類似度を測るための尺度としては、余弦尺度、内積尺度、ユークリッド距離尺度などを用いることができる。また、判別に関しては、ある類似度以上を削除対象として採用するような閾値処理による判別、類似度の高い類似度の高い一定数を削除対象として採用する定数処理、もしくは判別分析なども用いることができる。このようにして、採用された特徴次元を前記表現空間変換関数から削除することで前記表現空間変換関数を修正することができる。

文書データの分割

|| 先頭から200文字目の文字からその前後で最も近い句
|| 点までをひとつの分割文書とする。 ||

分割文書1

ユーザーの要因を反映するような文書分類をおこなうためのひとつの方針として、前記表現空間変換関数により構成される空間における不必要的特徴次元や、悪影響を及ぼすような特徴次元に対し削除や合成をおこなったり、逆にある特徴次元を強調させるための操作をすることが考えられる。しかし、前記表現空間変換関数により生成される空間の特徴次元は、前記文書解析部にて抽出される単語のうち意味的に似たものが複数結合したものと考えができるため、各特徴次元の意味的な解釈は極めて複雑かつ多義的なものであるため、ユーザーに各特徴次元の意味を提示することは極めて難しい。

分割文書2

そこで、ユーザーに分類に反映させたくない内容や強調したい内容をもつ文書や單語などの情報を指定させ、それらを前記表現空間変換関数により構成される空間に適切に射影し、それらと類似度の高い特徴次元や低い特徴次元を判別することで、操作をおこなう特徴次元を選択することが考えられる。ここでは、前記表現空間変換関数の特徴次元を操作する例として、ユーザーが指定するある文書と類似度の高い特徴次元の削除を行う例を示す。

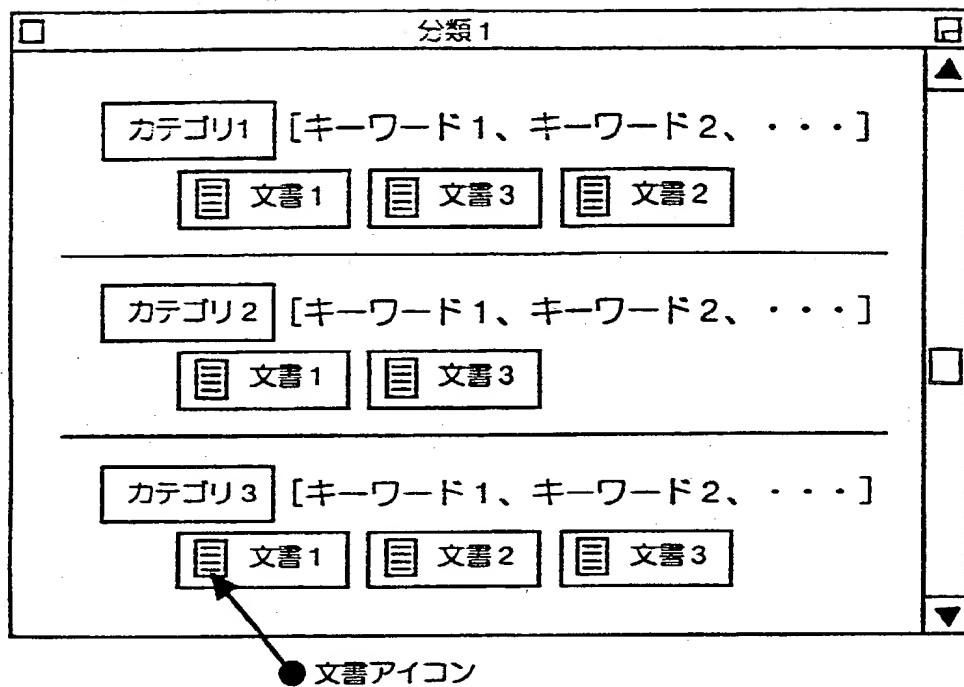
分割文書3

ユーザーにより指定された文書を前記文書特徴ベクトルと同じ次元数をもつベクトルで表現し、その文書ベクトルに前記表現空間変換関数を適用し文書ベクトルを前記表現空間変換関数により構成される空間へ射影する。そして、この射影された文書ベクトルと各特徴次元との類似度を算出することで、類似度の高い特徴次元を判別する。このとき、類似度を測るための尺度としては、余弦尺度、内積尺度、ユークリッド距離尺度などを用いることができる。

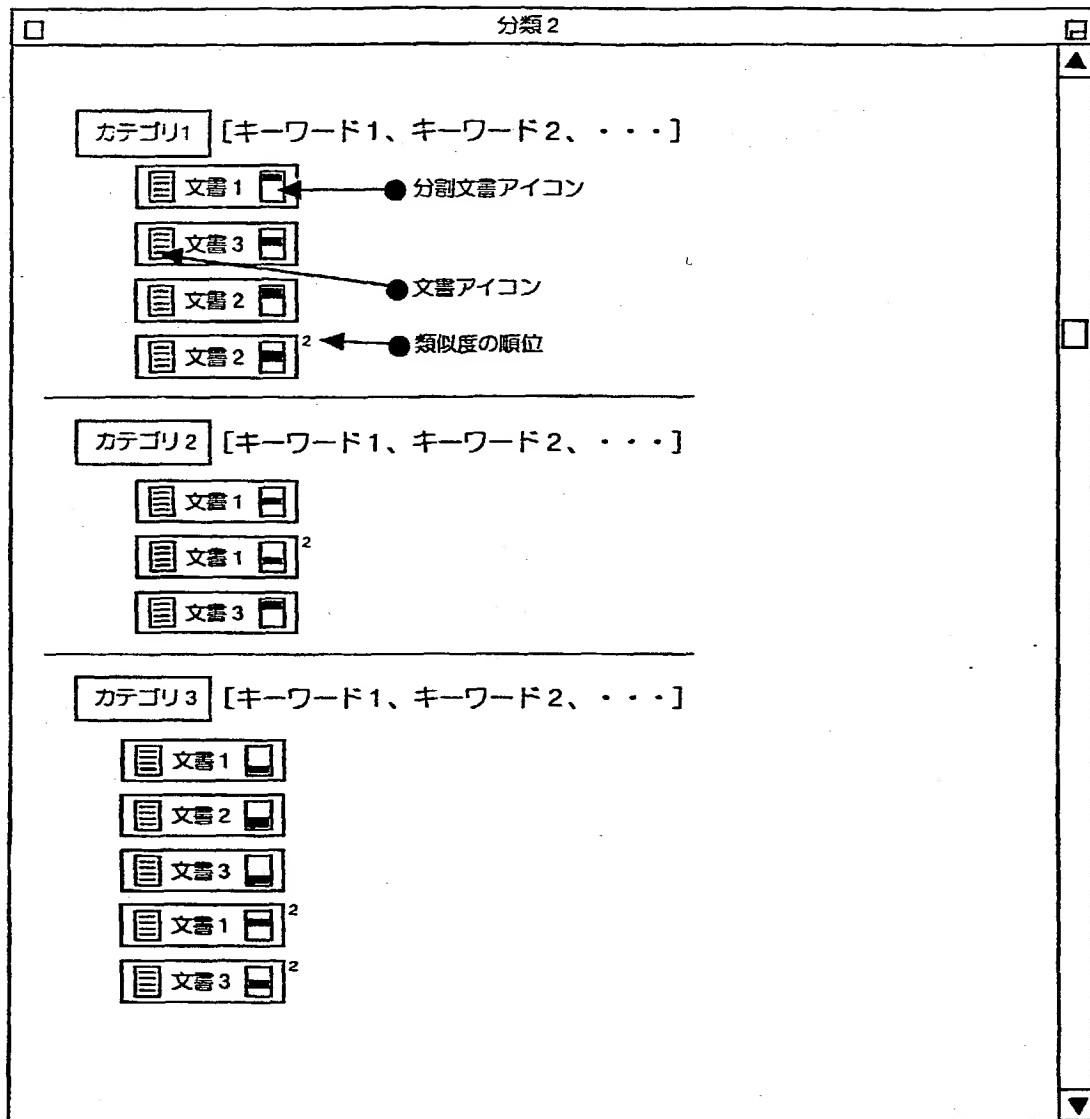
分割文書4

また、判別に関しては、ある類似度以上を削除対象として採用するような閾値処理による判別、類似度の高い類似度の高い一定数を削除対象として採用する定数処理、もしくは判別分析なども用いることができる。このようにして、採用された特徴次元を前記表現空間変換関数から削除することで前記表現空間変換関数を修正することができる。

【図55】



【図56】



【書類名】 要約書

【要約】

【課題】 文書の意味に係わるような分析作業において、単にその結果のみを出力するのではなく、情報分析作業全般にわたる支援をおこなうことを課題とする。

【解決手段】 入力された文書データを記憶する文書記憶部402と、文書記憶部402により記憶された文書データの全部または一部を選択する選択部403と、選択部403により選択された文書データの全部または一部から文字列の特徴に関するデータを抽出する特徴抽出部404と、特徴抽出部404により抽出された文字列の特徴に関するデータに基づいて文書データの全部または一部を加工処理する加工処理部405と、加工処理部405により加工処理された文書データの全部または一部を出力する出力部406とを備える。

【選択図】 図4

出願人履歴情報

識別番号 [000006747]

1. 変更年月日 1990年 8月24日

[変更理由] 新規登録

住 所 東京都大田区中馬込1丁目3番6号
氏 名 株式会社リコー